Dynamique des génomes et évolution du métabolisme lipidique chez les levures du clade Yarrowia

Stéphanie Michely

▶ **To cite this version:**

Stéphanie Michely. Dynamique des génomes et évolution du métabolisme lipidique chez les levures du clade Yarrowia. Génomique, Transcriptomique et Protéomique [q-bio.GN]. AgroParisTech, 2014. Français. NNT : 2014AGPT0021 . tel-01178514

HAL Id: tel-01178514

https://pastel.hal.science/tel-01178514

Submitted on 20 Jul 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Ecole Doctorale ABIES
Agriculture
Agriculture
Alimentation
Food
Biologie
Biology
Environnement
Environment
Santé
Health

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

L'Institut des Sciences et Industries du Vivant et de l'Environnement

(AgroParisTech)

Spécialité : Science de la vie et santé

présentée et soutenue publiquement par

Stéphanie MICHELY

le 19 mai 2014

Dynamique des génomes et évolution du métabolisme lipidique chez les levures du clade *Yarrowia*

Directeur de thèse : **Cécile Neuvéglise**

Jury

Mme Brigitte THOMASSET, DR CNRS, UMR 6022, Univ. de Technologie de Compiègne	Rapporteur
M. Olivier LESPINET, MC Paris XI, UMR 8621 Institut de Génétique et Microbiologie	Rapporteur
M. Thierry CHARDOT, DR INRA, UMR 1318 Institut Jean-Pierre Bourgin	Examinateur
M. Jean-Denis FAURE, Pr AgroParisTech, UMR 1318 Institut Jean-Pierre Bourgin	Examinateur
M. Emmanuel TALLA, MC Université de la Méditerranée, UMR 7283 LCB	Examinateur
Mme Cécile NEUVEGLISE, DR INRA, UMR1319 Institut Micalis	Directeur de thèse

AgroParisTech
UMR 1319 MICALIS
AgroParisTech 78850 Thiverval-Grignon

TABLE DES MATIERES

INTRODUCTION GENERALE 1

1 LES LEVURES, DES ORGANISMES MODELES EN GENOMIQUE 1
2 *YARROWIA LIPOLYTICA* 3
2.1 UN OUTIL PROMETTEUR 4
2.2 UN GENOME ORIGINAL 5
3 ET SI ON REGARDAIT LES ESPECES PROCHES ? 6
3.1 LE CLADE *YARROWIA* : PEU ETUDIE MAIS EN PLEIN ESSOR 6
3.2 POURQUOI S'INTERESSER AUX GENOMES DU CLADE *YARROWIA* ? 9

CHAPITRE 1: PHYSIOLOGIE COMPARATIVE DU CLADE YARROWIA 13

1 LES OBJECTIFS 13
2 STRATEGIE D'ANALYSE DES DONNEES DE CHROMATOGRAPHIE 14
3 « COMPARATIVE PHYSIOLOGY OF OLEAGINOUS SPECIES FROM THE YARROWIA CLADE » 15
4 CONCLUSION 15
ARTICLE 1 : COMPARATIVE PHYSIOLOGY OF OLEAGINOUS SPECIES FROM THE YARROWIA CLADE 18

CHAPITRE 2 : LE METABOLISME LIPIDIQUE DU CLADE YARROWIA 39

I. LES OBJECTIFS 39

II. INTRODUCTION 40

1 IMPORT DES LIPIDES 42
1.1 HYDROLYSE DES TRIGLYCERIDES : LES LIPASES 42
1.2 MULTIPLES VOIES METABOLIQUES POUR LES SUBSTRATS HYDROPHOBES 44
1.3 TRANSPORT ET ACTIVATION DES ACIDES GRAS DANS LE CYTOSOL 45
2 DEGRADATION DES ACIDES GRAS ACTIVES DANS LES PEROXYSOMES 46
2.1 ACTIVATION ET TRANSPORT DES ACIDES GRAS DANS LES PEROXYSOMES 46
2.2 BETA OXYDATION : UN CYCLE DE REACTIONS ET PLUSIEURS COMPARTIMENTS 48
3 L'ACETYL-COA 53
3.1 LA SYNTHESE DE L'ACETYL-COA 53
3.2 TRANSPORT DE L'ACETYL-COA 55
3.3 L'UTILISATION DE L'ACETYL-COA 57
4 LES METABOLISMES DES COFACTEURS : NADP ET NADPH 60
4.1 LA NAVETTE REDOX 61
4.2 LA NAVETTE CITRATE-OXOGLUTARATE DANS LA MITOCHONDRIE 62
4.3 LA NAVETTE MALATE/OXALOACETATE 62
5 SYNTHESE *DE NOVO* DES ACIDES GRAS 63
5.1 LE DECLENCHEMENT DE L'ACCUMULATION 63
5.2 LE CYCLE DE LA NEO SYNTHESE DES ACIDES GRAS 63
6 ACCUMULATION SOUS FORME DE LIPIDES NEUTRES 64
6.1 SYNTHESE DES TRIGLYCERIDES 64
6.2 SYNTHESE D'ESTERS DE STEROLS 66
6.3 LES CORPS LIPIDIQUES 66

III. RESULTATS 68
1 LES 4 NOUVEAUX GENOMES DU CLADE…ET CELUI DE OLHI 68
2 STRATEGIES 73
2.1 RECONSTRUCTION DU METABOLISME LIPIDIQUE 73
2.2 DEFINITION DES FAMILLES DE GENES 74
3 RECONSTRUCTION DU METABOLISME LIPIDIQUE 76
4 DE NOUVEAUX GENES POUR LE METABOLISME LIPIDIQUE 77
4.1 L'ALCOOL GRAS DESHYDROGENASE FAO 77
4.2 LES ACYL-COA SYNTHETASE FAA 78
4.3 L'ACETYL-COA SYNTHETASES ACS2 DANS LES PEROXYSOMES ? 79
4.4 L'ACYL-COA THIOESTERASE TES1 79
4.5 LE TRANSPORTEUR DE PYRUVATE MITOCHONDRIAL CODE PAR MPC1 ET MPC2 79

4.6 YALI A UN COMPLEXE BCKDC 80
4.7 LA 3-OXOACID COA-TRANSFERASE SKT1 81
4.8 UNE SECONDE TGL3 82
4.9 LA BETA OXYDATION MITOCHONDRIALE 83
4.10 LA NEO-SYNTHESE MITOCHONDRIALE DES AG 84
4.11 LES ENZYMES AUXILIAIRES 85
4.12 CAS D'EVOLUTION DE L'EPISSAGE ALTERNATIF 87

IV. CONCLUSION 91

CHAPITRE 3 : L'EVOLUTION DES GENOMES ET DU METABOLISME 95

I. LES OBJECTIFS 95

II. INTRODUCTION 96

1 EVOLUTION DES GENES ET SELECTION 96
1.1 LES SUBSTITUTIONS AFFECTANT UN GENE 97
1.2 L'ETUDE DE LA SELECTION 100
2 LA CREATION D'UNE FAMILLE DE GENES 102
2.1 LES DUPLICATIONS 102
2.2 LES MECANISMES A L'ORIGINE DE LA DUPLICATION GENIQUE 108
3 QU'EST CE QU'UNE FAMILLE DE GENES ? 113
3.1 HOMOLOGUE, ORTHOLOGUE, PARALOGUE OU OHNOLOGUES ? 113
3.2 LES GROUPES D'ORTHOLOGUES ET DE PARALOGUES 114
4 L'EVOLUTION DES FAMILLES DE GENES 115
4.1 LE PROCESSUS DE FIXATION NEUTRE 115
4.2 FIXATION PAR LA SELECTION POSITIVE 119
4.3 EVOLUTION DIVERGENTE, CONCERTÉE ET « BIRTH-AND-DEATH » 121
5 RETRACER L'HISTOIRE EVOLUTIVE DES FAMILLES DE GENES 123
5.1 IDENTIFICATION DES ORTHOLOGUES PAR HOMOLOGIE ET SYNTENIE 123
5.2 LA PHYLOGENIE D'UNE FAMILLE DE GENE 124
5.3 INCONGRUITE ENTRE ARBRE DE GENES ET ARBRE D'ESPECES OU ARBRES CONSENSUS 125

III. STRATEGIE 127

1 LA CONSERVATION DES PROTEINES 127
2 LES PRESSIONS DE SELECTION 128
3 RECONSTRUCTION DES SCENARIOS EVOLUTIFS DES FAMILLES DE GENES 129
4 STRATEGIE DES SCRIPTS 132
4.1 SCRIPT 1 : DN/DS 132
4.2 SCRIPT 2 : EVOLUTION DES FAMILLES DU METABOLISME LIPIDIQUE 134
4.3 SCRIPT 3 : LA SYNTENIE 135

VI. RESULTATS 137

1 VUE D'ENSEMBLE DE L'EVOLUTION DE CE METABOLISME 137
1.1 EXISTENT-ILS DES VOIES DU METABOLISME LIPIDIQUE QUI EVOLUENT DIFFEREMMENT ? 137
1.2 GENES DU METABOLISME LIPIDIQUE VS. GENOMES COMPLETS 139
2 EVOLUTION DES FAMILLES DE GENES 142
2.1 DES FAMILLES DYNAMIQUES EN EXPANSION DANS LE CLADE 142
2.2 VERS UNE CONTRACTION DES FAMILLES ? 145
2.3 DES FAMILLES SOUS PRESSION 148
3 ET PLUS PRECISEMENT ? 149
3.1 POX, UN EXEMPLE DE FAMILLE PEU DYNAMIQUE 149
3.2 LES EST, UN CAS D'ECOLE 154
3.3 LIP, UN EXEMPLE DE FAMILLE TRES DYNAMIQUE 156
ARTICLE 2 : EVOLUTIONARY GENOMICS: AN EFFICIENT TOOL TO EXPLORE ENZYME DIVERSITY AND GUIDE THEIR ENGINEERING **159**

V. CONCLUSION 207

DISCUSSION ET PERSPECTIVES 211

Introduction Générale

1 Les levures, des organismes modèles en génomique

Les levures sont des eucaryotes unicellulaires, appartenant au règne des Mycètes ou Mycota et au sous-règne des Dikarya, généralement de 3 à 4 µm de diamètre et présentant peu de différenciation cellulaire. Elles représentent un rameau précoce de la branche évolutive à l'origine des animaux. Elles sont présentes dans tous les biomes et sont très diversifiées (Starmer and Lachance, 2011). Elles se sont séparées des autres champignons il y au moins 300 à 400 millions d'années et se sont diversifiées en un grand nombre d'espèces. Il en existe environ 1500 connues mais les scientifiques considèrent que cela représente environ 10% de la totalité des espèces puisque de nombreuses espèces sont décrites chaque année. Les levures appartiennent à deux grands phylum, les basidiomycètes et les ascomycètes (Figure 0.1). Les basidiomycètes regroupent environ un tiers des espèces de dikarya et comprennent la majorité des champignons comestibles. Ils sont caractérisés par le fait que les méiospores sont externes. Le groupe des ascomycètes contient environ les deux tiers des espèces de dikarya décrites. Leurs méiospores sont produites à l'intérieur d'un sac ou asque. Ce groupe comprend par exemple les truffes mais aussi le sous-phylum des saccharomycotina (ou hémiascomycètes) avec les levures utilisées en boulangerie, brasserie et vinification. Les premières levures ascomycètes dont les génomes furent entièrement séquencés sont la levure boulangère *Saccharomyces cerevisiae* et la levure fissipare *Schizosaccharomyces pombe* (Goffeau et al., 1996; Wood et al., 2002).

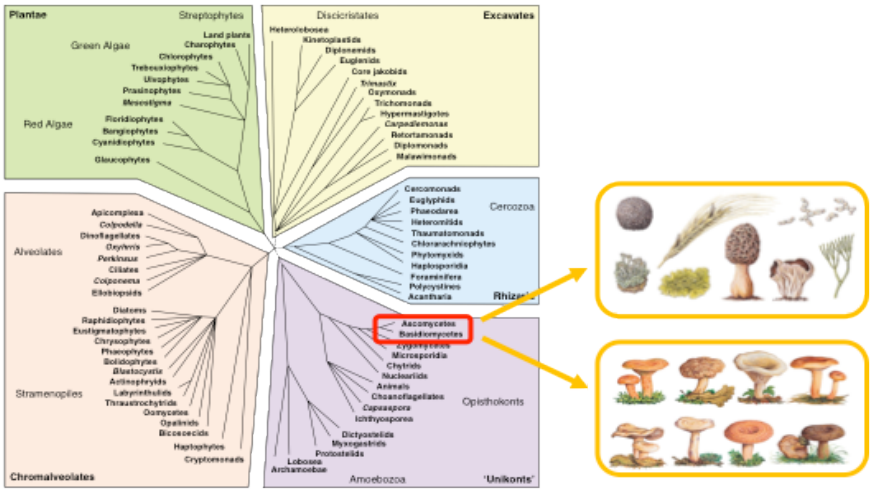


Figure 0.1 : Localisation des Ascomycètes et des Basidiomycètes sur l'arbre des eucaryotes de Keeling *et al.,* 2005 (Keeling et al., 2005).

Les saccharomycotina sont composés de plus de 1000 espèces différentes et ont divergé depuis 300 à 400 millions d'années (Kurtzman, 2011a). Elles ont colonisé tous les types de milieux grâce à leur formidable aptitude d'adaptation. La facilité à cultiver certaines de ces levures en laboratoire et à les manipuler génétiquement, associée à la compacité et la petite taille de leur génome, les ont propulsées au premier plan de la génomique. Le premier eucaryote séquencé, le génome de *Saccharomyces cerevisiae* (Goffeau et al., 1996), a rapidement été comparé à plus de 30 génomes de levures, dont environ un tiers descend de la duplication complète de l'ancêtre des *Saccharomyces* (Wolfe and Shields, 1997). La majorité des autres génomes séquencés appartient à des espèces considérées comme pathogènes avérés ou pathogènes émergents pour l'homme (Butler et al., 2009), mais finalement peu d'espèces environnementales. Ces génomes représentent une grande gamme évolutive, avec beaucoup de réarrangements chromosomiques et un fort degré de divergence des séquences entre certaines espèces (Dujon et al., 2004; Souciet et al., 2000). Mais ces génomes ne sont pas représentatifs de la phylogénie des levures car ils sont soit très éloignés soit extrêmement proches et de nombreux phylums sont inexplorés. Actuellement, une centaine d'espèces sont entièrement séquencées (Figure 0.2). Le projet Génolevures en particulier a permis le séquençage de huit espèces sélectionnées en fonction de leur intérêt en tant qu'agents pathogènes de l'homme ou de leur intérêt industriel ou environnemental, mais aussi en fonction de leur position phylogénétique. Les espèces actuellement séquencées par Génolevures sont réparties sur tout l'arbre phylogénétique des Saccharomycotina avec des levures dont l'ancêtre commun a subi la duplication complète de son génome, *S. cerevisiae* et *C. glabrata* ; des protoploïdes avec *Z. rouxii, L. thermotolerans, L. kluyveri, K. lactis* et *E. gossypii* ; des CTG avec *D. hansenii*, *Pichia sorbitophila* et *C. albicans* ; et *Arxula adeninivorans* et *Yarrowia lipolytica* qui se situent à la base des Saccharomycotina.

Les études comparatives des génomes ont permis de mettre en évidence des originalités qui ont accru nos connaissances en génomique comparative. Nous pouvons citer par exemple la présence remarquablement faible de duplications segmentales chez les protoploïdes ; la faible redondance de ces génomes avec 44% de paralogues chez *S. cerevisiae* contre au moins 65% chez les eucaryotes pluricellulaires; et la grande divergence intraspécifique des séquences chez *L. kluyveri* et *S. paradoxus* (Liti et al., 2009 ; The Génolevures Consortium et al., 2009 ; Schacherer et al., 2009). De même, la divergence interspécifique, entre *S. cerevisiae* et *S. uvarum*, avec 11% d'indels et 62% d'identité en nucléotides, est largement supérieure à celle observée entre l'homme et le chimpanzé (Chen and Li, 2001) et est similaire à la divergence entre l'humain et la souris, avec 12% d'indels et 66% d'identité en nucléotides (Kellis et al., 2003).

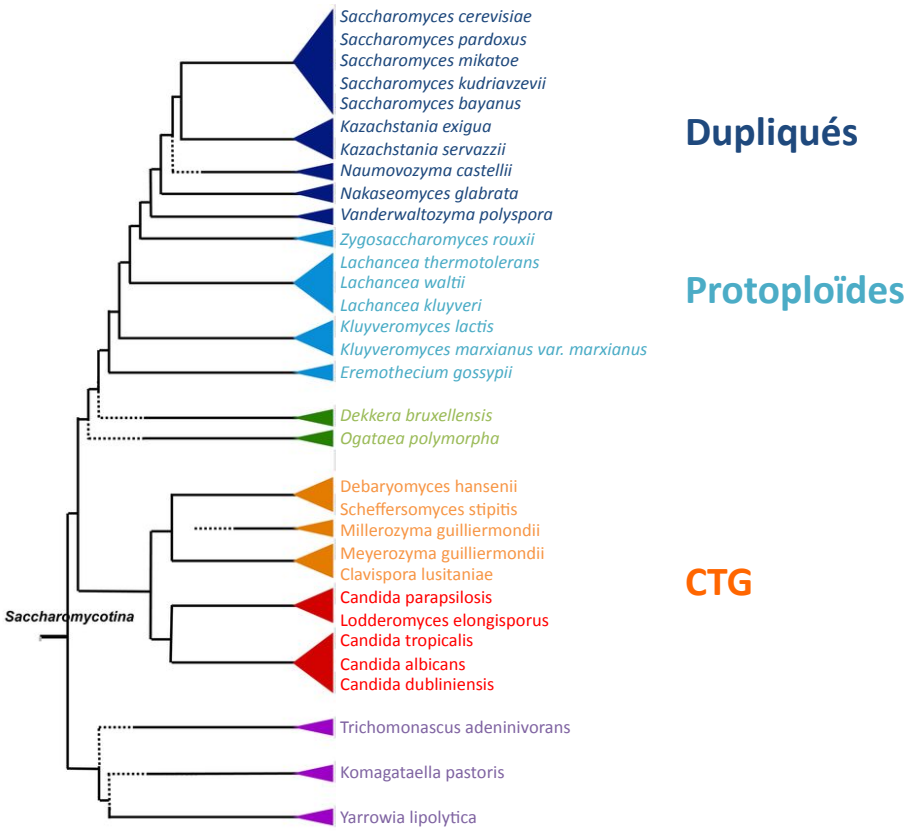


Figure 0.2 : Arbre des Hémiascomycètes, d'après Morales et Dujon (Morales and Dujon, 2012).

2 *YARROWIA LIPOLYTICA*

Yarrowia lipolytica (YALI), van der Walt et von Arx (1980), est l'espèce hémiascomycètes séquencée la plus distante de *S. cerevisiae* (SACE) (Dujon et al., 2004) et elle est très divergente des autres levures. Avec plus de 300 millions d'années qui la sépare de SACE, leurs différences sont autant au niveau physiologique, que métabolique, que pour l'organisation de leur génome. Malgré cela les études portant sur YALI utilisent souvent SACE comme comparaison car SACE est la levure la plus étudiée, avec de nombreuses fonctions de protéines validées expérimentalement donnant ainsi une référence.

YALI est un organisme saprophyte, vivant en aérobie stricte. La plupart des isolats sont haploïdes mais un état diploïde stable existe. YALI est connue pour ses activités lipolytiques et protéolytiques prononcées (Gardini et al., 2006; Ismail et al., 2001; Patrignani et al., 2011). YALI est ainsi principalement retrouvée dans les aliments avec une forte proportion de matières grasses et/ou de protéines, en particulier dans les produits laitiers et la viande. Son activité enzymatique lui permet de jouer un rôle important, dans l'affinage des fromages grâce aux composés aromatiques, aux alcools, aux lactones et aux ester qu'elle peut libérer (Suzzi et al., 2001). Elle est également isolée régulièrement dans des environnements naturels, souvent pollués, et dans de rares cas sur des patients en milieu hospitalier bien qu'elle ne soit pas

virulente. YALI est considérée comme levure non pathogène et a été classée comme organisme GRAS (Generally Recognized As Safe) par la Food and Drug Administration (FDA, États-Unis) (Groenewald et al., 2014). Elle est l'une des levures « non conventionnelle » les plus étudiées car elle est capable de produire des métabolites d'intérêt et a une grande capacité de sécrétion, ce qui justifie les nombreuses études réalisées dans le but de l'utiliser dans l'industrie, dans la biologie moléculaire et en génétique.

2.1 Un outil prometteur

YALI a la capacité de croitre sur de multiples substrats hydrophobes tels que les alcanes ou les acides gras (Nyns et al., 1968), sa dénomination même, « *lipolytica* », provient de sa capacité à hydrolyser les lipides. C'est une levure oléagineuse capable de stocker de grandes quantités de lipides, plus de 20% de son poids sec chez les souches sauvages et plus de 40% pour les souches améliorées (Beopoulos et al., 2011). Pour cela elle assimile des alcanes ou des acides gras ou des sucres et néo-synthétise des acides gras. Les acides gras sont alors dégradés dans les peroxysomes ou stockés dans les corps lipidiques sous forme de triglycérides.
YALI est également capable de métaboliser de façon très efficace une large gamme de substrats hydrophobes tels que des hydrocarbures, des acides gras, des huiles et graisses (Fickers et al., 2005a; Jain et al., 2004) . Ces substrats sont fortement énergétiques et peuvent être à l'origine de produits de transformation à grande valeur ajoutée dans le domaine de la production.
YALI doit ces capacités à des voies métaboliques particulières telles que la synthèse des acides gras, la ß-oxydation peroxysomale et mitochondriale, la synthèse d'acétyl-CoA à partir du citrate, l'assimilation et la dégradation des alcanes. Ces voies impliquent certaines protéines typiques des levures oléagineuses que sont par exemple l'ATP-citrate lyase, l'acyl-CoA déshydrogénase et l'utilisation de la 3-hydroxyacyl-CoA déshydrogénase pour la dégradation de la lysine. De plus, ces compétences sont facilitées chez YALI par la présence de familles contenant un grand nombre de gènes impliqués dans le métabolisme lipidique avec par exemple la famille des 16 lipases, celle des 10 cytochromes P450 et les 6 acyl-CoA oxydases. De même, une étude réalisée à partir des génomes de *S. cerevisiae*, *C. glabrata*, *K. lactis*, *D. hansenii* et YALI et portant sur l'évolution des transporteurs dans le phylum des hémiascomycètes, révèle un nombre particulièrement important de transporteurs peroxysomaux et de transporteurs d'acides gras chez *Yarrowia* par rapport aux autres levures (De Hertogh et al., 2006).

La première utilisation de YALI date des années 1960 pour la production de protéines (single cell protein production) avec des alcanes comme source de carbone. Depuis cette utilisation industrielle, de nombreuses nouvelles applications ont été testées dans les domaines de la bioconversion, de la production d'arôme, de la production de métabolites intermédiaires, dans

les processus de bioremédiation et dans la technologie alimentaire. Par exemple, YALI est capable, par oxydation des alcanes et des acides gras, de produire des acides dicarboxyliques DCA qui ont une grande importance industrielle pour la production de détergents, d'agents tensioactifs, de lubrifiants, de cosmétiques et de composés aromatiques. YALI est également l'organisme le plus prometteur, en terme de quantité, pour la production d'un composé aromatique naturel, la γ-décalactone, par bioconversion de l'huile de ricin (Schrader et al., 2004). YALI peut aussi être utilisée pour la production de lipides de composition prédéterminée, tel que le beurre de cacao, à partir d'un milieu de culture composé de résidus agro-industriels (Papanikolaou et al., 2003). YALI a de plus la particularité de produire et d'excréter dans le milieu une large gamme d'acides organiques, y compris des intermédiaires du cycle du TCA : les acides citrique et isocitrique, 2-cétoglutarique et pyruvique. Ces acides peuvent alors être produits à partir de sources de carbone de faible coût, comme des alcanes ou des huiles végétales simplement en limitant l'apport de certains composés durant la croissance. Des souches recombinantes de YALI sont capables d'accumuler plus de 7% de leur poids sec en Poly-Hydroxy-Alcanoates PHA (Haddouche et al., 2011). Cette capacité est extrêmement prometteuse car les PHA sont des polyesters ayant des propriétés thermoplastiques et élastomères utilisées dans les domaines agro-alimentaires. D'autres applications pour la fermentation industrielle concernent la fabrication de bioplastiques et de biocarburants, de médicaments et de bio-implants.

Durant les dernières décennies, YALI à su s'imposer comme usine cellulaire en biotechnologie ainsi qu'en tant que modèle pour l'étude fondamentale du métabolisme des substrats hydrophobes. C'est un modèle complet et très pratique qui offre une grande facilité pour sa manipulation génétique grâce à son haploïdie et dont toutes les approches en « -omics » sont facilitées par son génome entièrement séquencé et annoté (Casaregola et al., 2000).

2.2 Un génome original

Le génome complet de la souche E150 (CLIB122) de YALI a été séquencé, annoté puis publié en 2004 par le Consortium Génolevures (Dujon et al., 2004). Son génome se distingue de ceux des autres levures d'un point de vue structurel et fonctionnel. Son contenu en G+C est d'environ 49 %, ce qui est significativement plus grand que dans les autres génomes de levures.
Avec ses 20,5 Mb son génome, divisé en 6 chromosomes, est le plus grand des hémiascomycètes séquencés et publiés mais ses 6642 CDS en font un génome peu compact avec un gène pour 3,1 kb. En effet, il est une fois et demie plus grand que celui de *S. cerevisiae* mais ne contient que environ 800 gènes de plus. De même, il a environ 1300 gènes de plus que le génome de *K. lactis* qui est pourtant deux fois plus petit. Actuellement YALI a le génome hémiascomycète connu le plus riche en introns. Les modèles de gènes de YALI sont fractionnés en exon-intron dans 15%

des cas (soit 951 gènes), contre seulement 4% chez *S. cerevisiae*, et certains gènes peuvent contenir jusqu'à 5 introns. Ces introns ont une taille moyenne de 280 bp, allant de 41 bp jusqu'à 3478 bp. La présence de ces introns peut générer des transcrits alternatifs principalement par rétention d'intron et par épissage alternatif pour au moins 70 gènes (Mekouar et al., 2010). Ces introns peuvent également être corrélés à des départs de transcription alternatifs ou être positionnés en 5'UTR ou 3'UTR.

Il a également été mis en évidence la présence de mécanismes de contrôle des ARN, tels que la Nonsense-Mediated mRNA Decay (NMD) associée à l'existence d'un complexe de jonction d'exons (EJC). Ce dernier ayant disparu chez *S. cerevisiae*, suggère que l'évolution régressive qui touche les levures en général a été moins rapide chez cette espèce.

Cinq centromères, pour 6 chromosomes, ont été isolés, révélant des séquences CEN avec des motifs différents des CEN de *S. cerevisiae*, *K. lactis* et *C. albicans,* pouvant suggérer que YALI représente un stade précoce de l'évolution des régions centromériques. Les centromères de YALI ont une faible teneur en GC et sont de petite taille, ce qui permet de les manipuler aisément. Parmi les 6 sous unités codant le complexe de reconnaissance des origines de réplication, deux sont très divergentes de leur homologue chez *S. cerevisiae* (Barth et al., 2013).

Les gènes des sous unités de l'ADN ribosomique ont une organisation particulière. Les gènes codant les sous unités du 18S, 5.8S et 26S, à l'image de ceux de *Pichia pastoris*, sont regroupés en blocs au niveau de 6 régions subtélomériques du génome. Chez la majorité des espèces Saccharomycotina, l'ADNr est en cluster unique, interne au chromosome. L'organisation des gènes du 5S est aussi une originalité puisqu'en plus d'être dispersés dans tout le génome, environ la moitié des copies sont fusionnées à des gènes d'ARNt (Acker et al., 2008). De manière surprenante les gènes d'ARNt peuvent également être fusionnés aux RUF70, qui sont des ARN non codants de fonction inconnue.

Enfin, le faible niveau de similarité des gènes composant le génome de YALI avec ceux des autres levures en fait une levure hémiascomycète atypique à la frontière entre les levures hémiascomycètes et les champignons filamenteux ascomycètes.

3 Et si on regardait les espèces proches ?

3.1 Le clade *Yarrowia* : peu étudié mais en plein essor

Yarrowia lipolytica est le seul taxon connu du genre téléomorphe *Yarrowia*. Dans un premier temps, YALI a été identifiée comme anamorphe, et donc classée comme *Candida lipolytica.* En 1970, sa forme sexuée a été identifiée par Wickerham (Wickerham et al., 1970). Elle a été décrite dans un état haploïde hétérothallique, avec deux types cellulaires différents. Quand les souches appropriées de sexes opposés sont mélangées dans les conditions appropriées, elles sont

capables de se conjuguer et de produire un zygote diploïde se développant sous forme de mycélium et aboutissant à la formation de quatre ascospores haploïdes. YALI a alors d'abord été renommée *Endomycopsis lipolytica* (Wickerham et al., 1970), puis *Saccharomycopsis lipolytica* (Yarrow, 1972), avant d'être finalement classée sous le nom de *Yarrowia* lipolytica (Walt and Arx, 1980). Le genre *Yarrowia* correspond alors à l'état ascosporique de l'espèce *Candida lipolytica*. Les caractères phénotypiques définissant ce genre sont l'aspect de vrais hyphes, la morphologie cellulaire denticulée conidique, la variabilité de la morphologie des ascospores. Un autre caractère la définissant est la présence de coenzyme Q-9 dans le système de transport d'électron (Kurtzman, 2011b).

Comme nous avons vu précédemment, *Yarrowia lipolytica* est une espèce très distante des autres levures connues, y compris de certaines levures dégradant les *n*-alcanes, comme *Candida maltosa, Candida tropicalis, Debaryomyces hansenii* et *Pichia guilliermondii* (Fickers et al., 2005a). Elle est considérée comme une levure ascomycète « non-conventionnelle », à la frontière entre les levures hémiascomycètes et les champignons filamenteux ascomycètes. Cependant, des analyses phylogénétiques récentes établies à partir des séquences des domaines D1/D2 de l'ADNr 26S de certaines espèces ont montré que YALI partage un ancêtre commun avec des espèces qui lui sont phylogénétiquement proches. Ainsi, le clade *Yarrowia* a été présenté pour la première fois par Kurtzman, en 2005 (Kurtzman, 2005). A cette date il incluait les 6 espèces : *Aciculoconidium aculeatum, Candida bentonensis, C. hispaniensis, C. galli, Y. lipolytica et C. incommunis*. Le clade ainsi défini incluait des espèces relativement éloignées de YALI. A cette date, une autre espèce très proche de YALI, *C. deformans,* avait déjà été présentée par Bigey *et al.* comme un taxon différent de YALI mais elle été considérée comme synonyme à YALI par comparaison phénotypique par Kurtzman et Fell (Kurtzman and Fell, 1998). En 2007, Knutsen décrit trois nouvelles espèces proches de YALI et soutient l'appartenance de *C. deformans* à un taxon différent de YALI par l'analyse des gènes de D1/D2 de l'ADNr 26S. Il redéfinit alors le clade avec les 8 espèces suivantes : *Yarrowia lipolytica*, *C. alimentaria, C. deformans, C. galli, C. hispaniensis, C. hollandica, C. oslonensis* et *C. yakushimensis*. En 2008, l'espèce *C. phangngensis,* proche de YALI, a été décrite par Limtong *et al.* (Limtong et al., 2008).

Puis, en 2011, Kurtzman propose un nouveau groupe de 10 espèces pour le clade composé des 8 espèces proposées par Knutsen, *C. phangngensis* et *C. bentonensis* (Kurtzman, 2011b).

Au début de ce projet de thèse nous considérions que le clade *Yarrowia* était composé de *Yarrowia lipolytica*, *C. alimentaria, C. deformans, C. galli, C. hispaniensis, C. hollandica, C. oslonensis, C. phangngensis* et *C. yakushimensis*. *C. bentonensis*, dont la séquence du D1/D2 du rDNA ne s'aligne pas bien avec celle de YALI, avait été exclue du clade car considéré comme trop éloignée de YALI. Mais durant ce projet, nous avons décidé d'exclure également *C. hispaniensis* du clade au vu de sa divergence par rapport aux autres espèces du clade (voir Chapitre 2, III, 1).

Nous la considérons donc comme l'espèce de référence extérieure au clade et nous proposons comme nouvelle appellation : *Oleozyma hispaniensis* (OLHI) (Table 0.1).

Table 0.1 : Description et milieu d'isolement des souches du clade *Yarrowia* et de *O. hispaniensis*. Les abréviations utilisées dans ce manuscrit sont indiquées pour les souches étudiées au cours de ce projet de thèse.

Espèce	Numéro de souche	Abrévia-tion	Substrat d'isolement	Pays	Référence
Yarrowia lipolytica	CBS 7504 =W29	YALI	Egouts de Paris	France	(Barth and Gaillardin, 1996)
Yarrowia deformans	CBS 2071	YADE	Ongle	Australie	(Bigey et al., 2003)
Candida galli	CBS 9722	YAGA	Foie de poulet	Georgia, USA	(Péter et al., 2004)
Yarrowia yakushimensis	CBS 10253	YAYA	Estomac de termite	Japon	(Knutsen et al., 2007)
Candida oslonensis	CBS 10146	YAOS	Yaourt au kiwi	Norvège	(Knutsen et al., 2007)
Candida hollandica	CBS 4855	YAHO	Vache	Pays-Bas	(Knutsen et al., 2007)
Candida alimentaria	CBS 10151	YAAL	Jambon	Norvège	(Knutsen et al., 2007)
Candida phangngensis	CBS 10407	YAPH	Eau de mer	Thaïlande	(Limtong et al., 2008)
Yarrowia keelungensis	CBS 11062		Eau de mer	Thaïlande	(Chang et al., 2013)
Yarrowia divulgata	CBS 11013		Bacon	Danemark	(Nagy et al., 2013)
Yarrowia porcina	CBS 12935		Porc haché / Sédiments de rivière tropicale	Hongrie / Brésil	(Nagy et al., 2014)
Yarrowia bubula	CBS 12934		Bœuf haché	Hongrie	(Nagy et al., 2014)
Candida hispaniensis ou *Oleozyma hispaniensis*	CBS 9996 =Y-5580	OLHI (alias YAHI)	Larve de *Spondylus buprestoides*	Espagne	(Kurtzman, 2005)

Ce clade est alors un groupe monophylétique qui comprend 8 espèces décrites et d'origines très différentes : parfois isolées de l'environnement (eau de mer, sol), parfois dans des milieux complexes (intestin de termites), mais également de produits alimentaires (yaourts, charcuterie).

Jusqu'en 2012, toutes ces espèces n'étaient connues que pour leur caractérisation taxonomique. Elles portent encore le nom de genre *Candida*, à ne pas confondre avec les levures du genre *Candida* qui sont des pathogènes, car leur sporulation sexuée n'a jamais été observée (Lachance et al., 2011). Bien que ces levures soient proches de YALI, très peu de choses sont connues sur elles. Trois lipases ont été purifiées chez YADE et leurs gènes ont été clonés et séquencés (Bigey et al., 2003; Boutur et al., 1995; Muderhwa et al., 1985). Chez YAGA l'activité lipoxygénase

permettant la conversion d'isoeugénol en vanilline et acide vanillique a été caractérisée. En 2012, Gaillardin *et al.* ont étudié leurs génomes mitochondriaux . Ils ont ainsi fourni la preuve d'un degré élevé de conservation de l'ordre et du contenu en gènes et des séquences protéiques dans les génomes mitochondriaux entre les 6 espèces YALI, YAAL, YADE, YAGA, YAPH, confirmant ainsi que ces espèces sont en effet très étroitement liées. Ils révèlent également une plasticité élevée de ces génomes particulièrement concernant le contenu d'intron, le type de l'intron, et la taille de l'intron.

Les études portant sur ce clade semblent s'accélérer récemment. En effet, s'il a fallu plus de 5 ans pour définir le clade *Yarrowia* à 8 espèces, tel que nous le connaissons actuellement, 4 nouvelles espèces ont été décrites en seulement quelques mois : *Yarrowia divulgata et Yarrowia keelungensis* en fin 2013 et *Yarrowia porcina* et *Yarrowia bubula* en début 2014. Nagy *et al.* fournissent alors un arbre du clade basé sur les séquences des espaceurs internes transcrits (ITS) et des gènes D1/D2 ADNr 26S, présentant les 12 espèces du clade et OLHI qu'ils incluent dedans (Nagy et al., 2014) (Figure 0.3). De même, Groenewald et Smith ont décrit les formes téléomorphes de *C. deformans* et *C. yakushimensis*, leur donnant donc les noms *Yarrowia* deformans et *Yarrowia* yakushimensis (Groenewald and Smith, 2013).

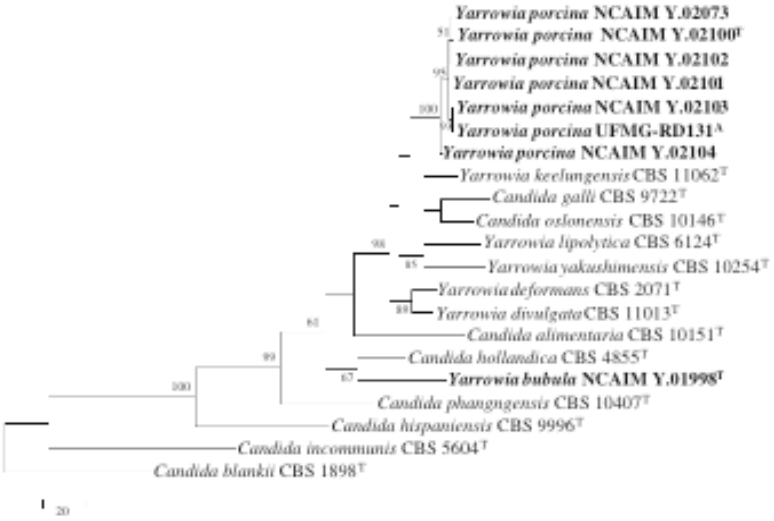


Figure 0.3 : Arbre phylogénétique proposé par Nagy *et al.*, basé sur les séquences concaténées des ITS et gènes D1/D2 ADNr 26S. *Candida blankii* a été utilisé en tant que groupe externe désigné. (Nagy et al., 2014).

3.2 Pourquoi s'intéresser aux génomes du clade *Yarrowia* ?

La diversité évolutive chez les levures est énorme, et les génomes disponibles sont soit très divergents (de *S. cerevisiae* à *Y. lipolytica*, environ 300 Millions d'années), soit appartenant à des espèces sœurs (les *Saccharomyces sensu stricto*, par exemple). Dans le premier cas, l'accumulation de centaines de réarrangements a abouti à une réorganisation massive de l'ordre des gènes et empêche toute reconstruction fiable des événements accumulés au cours de l'évolution. Dans le second cas, les génomes des espèces appartenant au clade des *sensu stricto*

sont extrêmement proches, ne différant que par quelques translocations (Fischer et al., 2000). Grâce aux nouvelles technologies de séquençage, il est maintenant possible de s'intéresser à l'évolution du génome de l'ensemble des espèces d'un clade afin de décrypter et mesurer l'impact des mécanismes d'évolution impliqués dans l'adaptation à leur milieu de vie. C'est ce qui à été fait pour les CTG et les Nakaseomyces par exemple (Butler et al., 2009; Gabaldon et al., 2013).

L'ensemble des originalités de YALI, au niveau métabolique comme génomique, font du clade *Yarrowia* un modèle d'étude sans précédent pour appréhender l'évolution des génomes.

PRÉSENTATION ET OBJECTIFS DE LA THÈSE

Dans ce contexte l'objectif de mon projet de thèse était de comprendre l'évolution du métabolisme lipidique au sein du clade *Yarrowia* et de déterminer les mécanismes évolutifs impliqués, en liant génomique comparative et génomique fonctionnelle lorsque cela est possible. Ce projet couvre alors deux aspects : l'analyse comparée du métabolisme lipidique des espèces du clade et l'analyse comparée des génomes du clade.

Les buts poursuivis étaient les suivants :

- L'établissement des profils phénotypiques des espèces du clade *Yarrowia* sur différents milieux de croissance afin d'étudier l'émergence du caractère oléagineux dans ce clade.
- La reconstruction du métabolisme lipidique chez *Y. lipolytica* et les autres espèces du clade en utilisant les données disponibles de génomique et de transcriptomique afin de savoir si leur métabolisme lipidique a les mêmes spécificités que celui de *Y. lipolytica.*
- La proposition de nouveaux gènes candidats pour les étapes du métabolisme lipidique non résolues.
- L'étude des profils transcriptomiques des espèces sur des milieux lipidiques (acide oléique, triglycérides) en s'intéressant au niveau d'expression des gènes impliqués dans le métabolisme des lipides.
- La caractérisation de l'évolution des gènes du métabolisme comparée à celle des génomes complets.
- La reconstruction des scénarios évolutifs des familles de gènes du clade, afin de mettre en évidence d'éventuelles expansions ou contractions de familles.

CHAPITRE 1 :

PHYSIOLOGIE COMPARATIVE DU CLADE *YARROWIA*

1 LES OBJECTIFS

Yarrowia lipolytica est capable de pousser sur de nombreux substrats hydrophobes tels que les alcanes ou acides gras. C'est une levure oléagineuse ayant de grandes capacités de synthèse et de stockage des acides gras. Ces caractéristiques font de cette levure un modèle pour l'étude du métabolisme lipidique.
Avant de commencer à s'intéresser à l'évolution du métabolisme lipidique dans le clade *Yarrowia*, nous avons décidé d'étudier l'émergence du caractère oléagineux dans ce clade. Cette étude est présentée ci-dessous sous forme d'un article publié dans PLOS ONE en 2013, Volume 8, Issue 5.

L'objectif de cette partie était de répondre à différentes questions :

- Sur quels milieux poussent les différentes espèces du clade et OLHI ?
- Ont-elles les mêmes caractéristiques que YALI ?
- Quelles sont leurs capacités de stockage et de synthèse des acides gras ?
- Sont–elles toutes oléagineuses ?

Lorsque cette étude a été réalisée nous disposions de souches sauvages de 8 espèces du clade : YALI, YAGA, YAYA, YADE, YAPH, YAHO, YAOS, YAAL et de l'espèce externe au clade OLHI. Les données physiologiques disponibles pour ces espèces sont aujourd'hui encore limitées, et proviennent principalement des tests d'assimilation effectués lors de leur identification. Il avait déjà été montré que toutes ces espèces présentent une diversité phénotypique (Kurtzman, 2011).

Nous avons alors étudié les capacités de 9 souches à utiliser divers substrats hydrophobes ou non hydrophobes comme principale source de carbone. Ces souches étudiées correspondent aux souches types de chacune des 9 espèces, sauf pour YALI, pour laquelle nous avons pris la souche W29, car la souche type 6124 est diploïde. Nous avons également établi les cinétiques de

croissance sur glucose et sur acide oléique. Puis nous avons défini leurs profils lipidiques, ainsi que leur capacité à accumuler et à synthétiser des lipides.

2 Stratégie d'analyse des données de chromatographie

Pour tester les capacités de stockage et de synthèse des acides gras nous avons dosé, par chromatographie en phase gazeuse (GC), la quantité de lipides présents dans les cellules des différentes souches ayant poussé soit sur glucose, soit sur acide oléique à différents stades physiologiques. Afin d'éviter l'analyse manuelle d'un grand nombre de données j'ai écrit un script en Python et R (Figure 1.1).

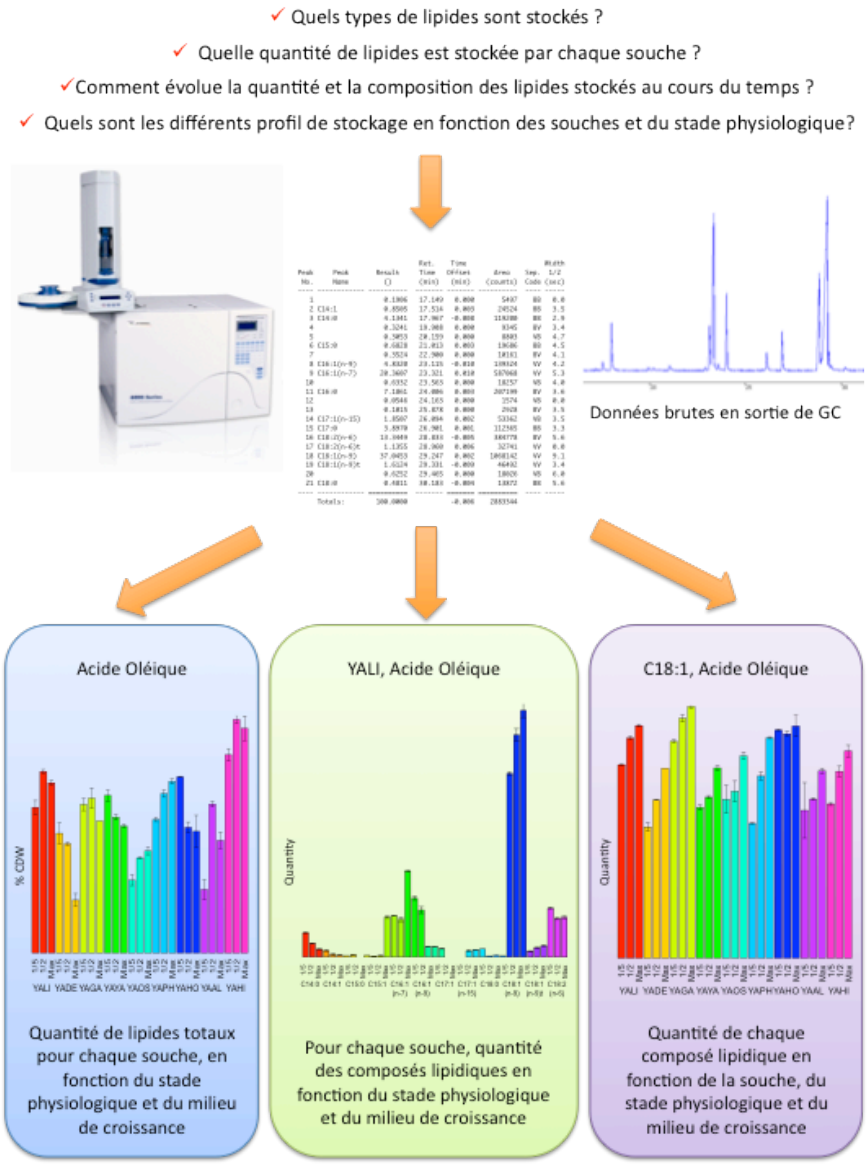


Figure 1.1 : Stratégie du script permettant d'analyser les résultats bruts de dosage des lipides par chromatographie en phase gazeuse.

Ce script prend en entrée les données brutes sorties de la GC et peut accepter différents types de conditions à comparer telles que le stade physiologique, le milieu, l'espèce, la souche (mutante ou non)… Dans le cas de cette étude nous avons comparé les données obtenues en fonction du milieu, en fonction de la souche et en fonction de la phase de croissance.

Le script génère alors des tableaux et des graphiques permettant de comparer les contenus en lipides selon les conditions étudiées. Dans notre cas il génère trois types de graphiques :
- La quantité de lipides totaux pour toutes les souches aux différents stades physiologiques, donnée en microgramme ou en pourcentage du poids sec de la cellule (% CDW) lorsque que le poids est renseigné. Ces graphiques permettent de comparer la capacité globale de synthèse et/ou de stockage des souches et leur profil en fonction du temps.
- Le type et la quantité de chacun des composés stockés et/ou synthétisés pour une souche, en fonction du stade physiologique. Cela permet de repérer, pour chaque stade physiologique, les composés qui sont accumulés à partir du milieu, ceux qui sont utilisés par la cellule, ceux qui sont stockés et ceux qui sont néo synthétisés. Nous pouvons donc en déduire les cinétiques de stockage et de remobilisation de chaque souche.
- Le troisième type de graphique généré représente pour un composé présent sa quantité en fonction de la souche et du temps. Ces graphiques sont surtout utiles dans le cas des études de différences entre une souche sauvage et un mutant que l'on cherche à améliorer pour le stockage et/ou la synthèse d'un composé en particulier.

3 « Comparative Physiology of Oleaginous Species from the *Yarrowia* Clade »

Cette étude de physiologie comparative a donnée lieu à la publication d'un article scientifique présenté à la fin de ce Chapitre, pages 17 à 25. Pour éviter toute confusion, il est important de noter qu'au moment de la rédaction de ce papier nous considérions OLHI comme espèce appartenant au clade *Yarrowia* et nous l'appelions encore *Yarrowia hispaniensis*, donc abrégé YAHI (voir Chapitre 2, III, 1).

4 Conclusion

Le premier aspect de ce projet de thèse consistait à déterminer la physiologie de 9 espèces du clade *Yarrowia* afin de répondre à plusieurs questions : sur quels milieux poussent ces espèces ? Ont-elles les mêmes capacités de stockage et de synthèse des lipides que YALI ? Et ont-elles les

mêmes caractéristiques que YALI ? Pour répondre à ces questions nous avons utilisé diverses techniques de microbiologie et de biochimie telles que les tests de croissance en milieu liquide et solide, l'établissement de cinétique de croissance et l'extraction et le dosage des lipides par chromatographie en phase gazeuse.

Pour les tests de cultures nous avons utilisé 2 types de substrats : des substrats hydrophobes et des sucres. Les substrats hydrophobes mettent en jeu chez YALI les capacités de stockage, de synthèse et de dégradation des lipides tandis que les sucres mettent en jeu les capacités de synthèse *de novo* des lipides. Nous avons alors fait des tests en gouttes sur composés hydrophobes, de types lipides et alcanes, allant du C4 au C22 et nous avons fait des tests de croissance sur sucres en utilisant des galeries API et des cultures en microplaques. Nous avons alors testé 44 milieux différents de type sucres, substrats hydrophobes et non hydrophobes. Nous avons plus particulièrement étudié leur croissance sur glucose et acide oléique. Ceci nous a permis de déterminer leur capacité de stockage et de synthèse des acides gras à 3 étapes différentes de la phase exponentielle de croissance ainsi le type de composés lipidiques stockés et synthétisés. Pour toutes ces études nous avons rencontré un problème de reproductibilité des manipulations avec les espèces autres que YALI, provenant probablement du fait que ce sont des souches sauvages contrairement à YALI qui est probablement maintenant une souche domestiquée. Il aurait alors été plus rigoureux d'augmenter le nombre de réplicats biologiques. D'autre part nous avons testé une seule concentration de substrat pour chaque milieu de culture alors que nous savons maintenant que les réponses des souches varient radicalement en fonction de la quantité de substrats. Et enfin nous n'avons testé qu'une souche par espèce. Sachant que chaque souche d'une même espèce peut être isolée de milieux et de géographies très variables, cette étude ne nous permet pas de généraliser nos résultats à l'ensemble de l'espèce. Par exemple, nous avons observé une incapacité de la souche française W29 de YALI à pousser sur fructose à la concentration testée tandis que d'autres personnes de notre équipe ont montré que la souche allemande H222 de YALI pousse très bien sur ce sucre.
Les résultats obtenus dans cette étude de physiologie comparative nous ont permis de conclure que toutes les souches testées, comme YALI, sont capables de pousser sur une large gamme de milieux alcanes et acides gras avec des réponses légèrement différentes en fonction du type de composé et de la longueur des chaines carbonées des substrats. Par exemple, YAPH et OLHI ont une meilleure capacité à croitre sur acide oléique. Par contre, leur utilisation des substrats non hydrophobes comme seule source de carbone est très limitée car, parmi tous les sucres testés, elles ne sont capables de croitre que sur les milieux glucose, fructose et glycérol. Il est également intéressant de remarquer que leurs capacités à pousser sur les différents substrats ne semblent pas être liés à leur phylogénie. Ceci suggère qu'il y a des pertes ou gains de fonction espèce spécifique, soit au niveau de leur contenu en gène soit au niveau de leur régulation des gènes.

Un des deux points majeurs de cette étude est l'aptitude de OLHI, l'espèce out-group du clade, à pousser sur un plus large panel de substrats ainsi que sa capacité à stocker jusqu'à 30% de lipides de plus que YALI dans nos conditions. Cette étude place alors OLHI comme un model prometteur dans le milieu des biotechnologies mais aussi comme un réservoir de gènes intéressants pour l'amélioration de YALI. Le second point fort est la capacité de stocker au moins 30% du poids sec en lipides observée dans toutes les souches du clade testées. Cette observation nous permet de caractériser toutes ces souches comme des levures oléagineuses.

L'étude comparative de la physiologie des espèces du clade *Yarrowia* et OLHI, nous a donc permis de mettre en évidence quelques différences entre les souches quant à leur utilisation des substrats et à leurs capacités de stockage et de synthèse des acides gras. La question qui se pose alors est : quelles sont les bases génétiques de ces différences physiologiques ? Pour la suite de mon projet de thèse nous avons donc utiliser des approches de génomique comparative et de transcriptomique pour répondre en partie a cette question.

OPEN ACCESS Freely available online

PLOS ONE

Comparative Physiology of Oleaginous Species from the *Yarrowia* Clade

Stéphanie Michely[1], Claude Gaillardin[2], Jean-Marc Nicaud[3], Cécile Neuvéglise[1]*

1 INRA, UMR 1319 Micalis, Jouy-en-Josas, France, **2** AgroParisTech, Micalis, Jouy-en-Josas, France, **3** CNRS, Micalis, Jouy-en-Josas, France

Abstract

Yarrowia lipolytica is a genetically tractable yeast species that has become an attractive model for analyses of lipid metabolism, due to its oleaginous nature. We investigated the regulation and evolution of lipid metabolism in non-Saccharomycetaceae yeasts, by carrying out a comparative physiological analysis of eight species recently assigned to the *Yarrowia* clade: *Candida alimentaria, Y. deformans, C. galli, C. hispaniensis, C. hollandica, C. oslonensis, C. phangngensis* and *Y. yakushimensis*. We compared the abilities of type strains of these species to grow on 31 non hydrophobic (sugars and other carbohydrate compounds) and 13 hydrophobic (triglycerides, alkanes and free fatty acids) carbon sources. Limited phenotypic diversity was observed in terms of the range of substrates used and, in the case of short-chain fatty acids, their toxicity. We assessed the oleaginous nature of these species, by evaluating their ability to store and to synthesize lipids. The mean lipid content of cells grown on oleic acid differed considerably between species, ranging from 30% of cell dry weight in *C. oslonensis* to 67% in *C. hispaniensis*. Lipid synthesis in cells grown on glucose resulted in the accumulation of C18:1 (n-9) as the major compound in most species, except for *C. alimentaria* and *Y. yakushimensis*, which accumulated principally C18:2(n-6), and *C. hispaniensis*, which accumulated both C16:0 and C18:1(n-9). Thus, all species of the clade were oleaginous, but they presented specific patterns of growth, lipid synthesis and storage, and therefore constitute good models for the comparative analysis of lipid metabolism in this basal yeast clade.

Citation: Michely S, Gaillardin C, Nicaud J-M, Neuvéglise C (2013) Comparative Physiology of Oleaginous Species from the *Yarrowia* Clade. PLoS ONE 8(5): e63356. doi:10.1371/journal.pone.0063356

Editor: Steven Harris, University of Nebraska, United States of America

Received January 15, 2013; **Accepted** April 1, 2013; **Published** May 7, 2013

Copyright: © 2013 Michely et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: SM has a PhD fellowship funded by the French Minister of Research via the ABIES doctoral school. This work was supported in part by the Institut National de la Recherche Agronomique (INRA) and additionally by the SAS PIVERT, in the frame of the French Institute of Excellence in the field of Low-Carbon Energies (IEED) P.I.V.E.R.T (www.institut-pivert.com) selected as an Investment for the Future ("Investissements d'Avenir"). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Cecile.Neuveglise@grignon.inra.fr

Introduction

Some yeast species can store and synthesize lipids from different carbon sources. Yeast species are described as "oleaginous" if the lipids they accumulate account for more than 20- to 25% of their biomass [1]. Oleaginous yeasts are dispersed throughout the entire phylogenetic tree of basidiomycetes and hemiascomycetes. The hemiascomycetous yeast *Yarrowia lipolytica* has emerged as an important model for lipid metabolism studies. Like other oleaginous yeasts, it can grow on sugars, such as glucose [2,3], and on hydrophobic substrates (HS) [4]. It can also synthesize and store lipids [5]. In addition, *Y. lipolytica* is highly tractable genetically, making it a good model species for biotechnological applications, particularly for single-cell oil production [6,7,8]. However, the amount of lipid that accumulates depends on the strain, the carbon source and growth conditions. Under optimal conditions, some wild strains of *Y. lipolytica* can store 36 % of their cell dry weight (CDW) as lipids [1]; similar levels are observed in fed-batch cultures with glucose/glycerol [9]; 43% of the CDW may be lipid in continuous fermentations of industrial glycerol [10] and up to 54% may be lipid in batch cultures on a stearin-based medium [11,12]. However, in flask cultures in which nitrogen concentration is not controlled, wild strains of *Y. lipolytica* do not generally accumulate more than 15% of their CDW as lipids when grown in glucose medium [13,14,15] or in wastewater [14,15].

Most of the lipids accumulating in *Y. lipolytica* are triacylglycerols rather than free fatty acids (FFA), the ratio of these two types of compounds being 5/1 (triacylglycerols/FFA) [16]. C16 and C18 compounds are the most abundant lipids stored by this yeast. However, their relative quantities depend on the growth medium used. The *Y. lipolytica* strain W29 (=CBS 7405) stores mostly C18:1 (54%), C16:0 (26%), C18:2 (12%) and a little C16:1 when cultured on glucose, whereas it accumulates C18:1 (66%), C16:1 (16%), C18:2 (9%) and a little C16:0 when cultured on oleic acid [17].

We investigated the emergence of oleaginous properties in yeasts, by comparative studies of *Y. lipolytica* W29 and strains from the eight species recently identified as members of the *Yarrowia* clade: *Candida (C.) alimentaria, Y. deformans, C. galli, C. hispaniensis, C. hollandica, C. oslonensis, C. phangngensis* and *Y. yakushimensis* [18,19,20,21,22,23]. These strains were isolated from various biotic and abiotic environments. Limited physiological data for these strains are available, mostly based on the assimilation tests used for their identification. All these strains are non fermentative, grow as a mixture of cells and hyphae and use a very limited range of sugars, mostly glucose, as carbon sources. Additional data for growth on other carbon sources have provided evidence of phenotypic diversity [24], although different tests were carried out on different strains. For example, *C. hispaniensis* and *C. oslonensis* have been reported to use galactose and sorbose, which are only

weakly metabolized, if at all, by other species, whereas *C. hispaniensis* is the only one of the species considered able to make use of trehalose. *C. galli* is the only one of these species that has been reported to grow in a vitamin-free environment; the failure of the other species to do so may result from thiamin auxotrophy, as reported in *Y. lipolytica* [25]. Tolerance to 10% NaCl differs between species, as does maximum growth temperature, which ranges from 27°C for *C. alimentaria* to 37°C for *C. phangngensis*, with most strains of other species growing little at temperatures above 30 to 32°C. It remains unclear whether this phenotypic diversity extends to the metabolism of hydrophobic compounds and, more generally, whether the oleaginous character of *Y. lipolytica* is particular to this species or common to some or all members of its clade. For example, data for growth on hexadecane and lipase production are patchy or absent for these species, with the exception of *Y. deformans*, in which three lipases have been purified and the corresponding genes have been cloned and sequenced [26,27,28].

We thus investigated the capacities of the type strains of the eight species and of the French wild-type strain W29 to use various hydrophobic and non hydrophobic substrates as carbon sources. Detailed growth kinetics were obtained on glucose and oleic acid media. Lipid profiles were determined, together with the capacity to accumulate and to synthesize lipids. The knowledge about the physiology of these strains provided by this study opens up new possibilities for genetic and transcriptomic analyses within the *Y. lipolytica* clade. The long-term objective will be to obtain a full understanding of lipid metabolism in this group, to improve the suitability of *Y. lipolytica* as a tool for biotechnological applications.

Materials and Methods

Yeast strains, media and growth conditions

The strains of the *Yarrowia* clade investigated in this study, their origins and references are listed in Table 1. All are wild-type prototroph strains. Strain names are abbreviated as follows: *YALI* (*Y. lipolytica* W29), *YAYA* (*Y. yakushimensis* CBS10253), *YADE* (*Y. deformans* CBS2071), *YAGA* (*C. galli* CBS9722), *YAOS* (*C. oslonensis* CBS10146), *YAHO* (*C. hollandica* CBS4855), *YAPH* (*C. phangngensis* CBS10407), *YAAL* (*C. alimentaria* CBS10151), *YAHI* (*C. hispaniensis* CBS9996). All strains were cultured at 28°C, with the exception of *YAAL*, for which the optimal growth temperature was 21°C (Additional Figure S1). Minimal medium base (MMB) contained 0.17% (wt/vol) yeast nitrogen base without amino acids and ammonium sulfate (Difco, Paris, France), 0.5% (wt/vol) NH_4Cl and 50 mM phosphate buffer pH 6.8; 1.5% (wt/vol) agar was added when necessary. The solid rich medium base (YP) contained 1% (wt/vol) yeast extract and 1% (wt/vol) peptone, together with 1.5% (wt/vol) agar. Carbon sources were added at a final concentration of 2%; hydrophobic substrates were first emulsified by sonication of a 20% mixture in the presence of 0.625% Tween 40.

Growth tests

Drop tests were performed with the 13 hydrophobic substrates (HS) listed in Table 2. Both solid rich medium and solid minimal medium were supplemented with 2% emulsified hydrophobic substrates, with the exception of alkanes, for which a paper inserted into the plate lid was soaked with alkane daily, to offset the effects of evaporation. Precultures were grown on plates of minimal medium containing 0.17% (wt/vol) yeast nitrogen base without amino acids and ammonium sulfate (Difco, Paris, France), 0.5% (wt/vol) $(NH_4)_2SO_4$ and 1% glucose. We plated 3 µl of each of a set of five-fold dilutions, corresponding to 2 to 1250 cells. Pictures were taken daily or every two days over a period of three weeks. Growth was considered to be delayed, weak or slow with respect to the strain with the best growth in the study [29]. API ID 32 C galleries (Biomérieux, France) of 32 cupules (31 different carbon substrates plus one control), were used to evaluate the assimilation of a set of carbon sources (Additional Table S1).

Microtiter plate culture analysis

Growth tests were performed in 100 µl cultures in 96-well plates, with constant shaking, in the presence of 2% arabinose, fructose, glucose, N-acetyl-glucosamine, ribose, saccharose or xylose as the carbon source. Precultures were grown on minimal medium plates, as for the growth tests. Growth was monitored by measuring the optical density at 600 nm (OD_{600} nm) at 10-minute intervals, with a microtiter plate reader (Biotek, Colmar, France). For each strain and set of conditions, we used two biological replicates.

Growth curves and parameter determination

The nine strains were grown in MMB plus 0.15% yeast extract, with 2% glucose or 2% oleic acid, in 500-ml Erlenmeyer baffled flasks, to improve the dispersion of alkanes and oxygen supply [30]. Cells from overnight YPD cultures were used to inoculate the culture at an initial OD_{600} of 0.5. Biomass production was followed by measuring OD_{600nm} every two or three hours. Cells grown in the presence of oleic acid were washed twice with 0.5% bovine serum albumin and then once with 0.9% NaCl before OD determination.

We used a custom-developed R software script [31] to derive parameters from the growth curves: length of the lag phase λ, growth rate corresponding to the maximum slope (µmax) and maximum cell density.

Lipid analysis

For fatty acid (FA) determinations, the strains were precultured and cultured under the same conditions and in the same media as for growth curve experiments. Two biological replicates were carried out. Samples of yeast cells were taken at three different growth stages and centrifuged at 3000 g for 5 min. The cell pellets were washed once with a half volume of 0.9% NaCl for glucose conditions and twice with a half volume of 0.5% bovine serum albumin and then once with a half volume of 0.9% NaCl for oleic acid conditions. Washed cells were freeze-dried for 24 h at 80°C.

The FA from the equivalent of 10 OD_{600} units of freeze-dried harvested cells were converted into their methyl esters by the Browse method [32]. The FA methyl esters were analyzed by gas chromatography, with a Varian 430-GC instrument equipped with a flame ionization detector and a CP SIL SCB low Bleed/MS column, for which the bleed specification at 260°C was 3 pA (30 m, 0.25 mm, 0,25 µm). FA were identified by comparison with commercial FA methyl ester standards (FAME32; Supelco) and quantified by the internal standard method, involving the addition of 75 µg of commercial C17:0 (Sigma). We determined the FA compositions from chromatograms with custom-developed Python and R software scripts [31].

Phylogeny

The phylogenetic tree was constructed by the concatenation of seven markers known to be representative of species evolution in fungi [33]. These markers were MS277 (YALI0B08756g homologs), MS444 (YALI0A00264g homologs), MS456 (YALI0B18722g homologs), MS561 (YALI0B16434g homologs), FG598 (YALI0A02695g homologs), and FG610 (YALI0D11220g

Table 1. Characteristics of the strains used in this study.

Species	Strain number	Abbreviation	Place of isolation	Country	Reference
Yarrowia lipolytica	CBS 7504 =W29	*YALI*	Parisian sewer	France	[25]
Yarrowia deformans	CBS 2071	*YADE*	Fingernail	Austria	[23]
Candida galli	CBS 9722	*YAGA*	Chicken liver	Georgia, USA	[18]
Yarrowia yakushimensis	CBS 10253	*YAYA*	Termite gut	Japan	[23]
Candida oslonensis	CBS 10146	*YAOS*	Kiwifruit yogurt	Norway	[20]
Candida phangngensis	CBS 10407	*YAPH*	Seawater	Thailand	[21]
Candida hollandica	CBS 4855	*YAHO*	Back of a cow	The Netherlands	[20]
Candida alimentaria	CBS 10151	*YAAL*	Cured ham	Norway	[20]
Candida hispaniensis	CBS 9996 = Y−5580	*YAHI*	*Spondylus buprestoides* larva	Spain	[19]

Abbreviations: CBS, Centraalbureau voor Schimmelcultures, Utrecht, The Netherlands; =W29, other common name used; =Y−5580, name from the NRRL collection.
doi:10.1371/journal.pone.0063356.t001

homologs), together with EF-1alpha (YALI0C09141g homologs). DNA was extracted as described by Hoffman and Winston [34]. PCR amplifications were performed with an Eppendorf 2720 thermocycler, ex-*Taq* polymerase (Takara) and the primers listed in Additional Table S2. Both strands were sequenced by GATC Biotech (Mulhouse, France).

The Staden package was used to analyze sequencing reads [35]. We used MUSCLE [36] to align sequences, and columns of gap-containing residues were removed manually. A final alignment of 2204 amino acids was obtained after sequence concatenation. Phylogenetic trees were constructed either by neighbor-Joining in ClustalX [37], or by maximum likelihood, with PHYML [38] and a JTT substitution model corrected for heterogeneity between sites by a Γ-law distribution, with four different categories of evolution rates. The proportion of invariable sites and the α-parameter of the Γ-law distribution were optimized according to the data. A bootstrap value was calculated from 100 replicates.

Results

Use of non hydrophobic carbon sources

Carbon assimilation tests have been performed for identification of the different species of the *Yarrowia* clade [18,19,20,21,22,23]. However, to obtain a homogeneous set of data, we assessed the ability of the nine strains to use 31 compounds. API galleries were initially used, for 30 compounds. The results were compared with the data available from the CBS (http://www.cbs.knaw.nl/). For 11 carbon sources, our observations were not consistent with those of the CBS for some strains, possibly reflecting differences in experimental procedures (API galleries, microtiter plate or agar plate tests) or carbon source concentration. Thus, liquid cultures in microtiter plates were used for these substrates, with glucose and sucrose used as positive and negative controls, respectively. These experiments confirmed that *YALI* and *YAPH* did not grow on arabinose, that *YAGA* grew on glycerol, and that *YALI* and *YADE* grew on N-acetyl-glucosamide. We also tested fructose as a carbon source under the same conditions (Additional Table S1).

Finally, only three of the 31 compounds (fructose, glycerol and glucose) supported the growth of all species and none of the species

Table 2. List of hydrophobic substrates used.

Class of hydrophobic substrate	Substrate	Number of carbon atoms and insaturation number
Triglycerides	tributyrin	C4
	triolein	C18:1
Methylates	methyl hexanoate	C6
	methyl decanoate	C10
	methyl myristate	C14
	methyl palmitate	C16
	methyl oleate	C18:1
Free fatty acids	hexanoic acid	C6
	oleic acid	C18:1
	erucic acid	C22:1
Alkanes	decane	C10
	dodecane	C12
	hexadecane	C16

doi:10.1371/journal.pone.0063356.t002

could grow on 21 other carbon sources. For the remaining seven compounds, differences in growth were observed between species. *YAHI*, which belongs to the most divergent species, was unique in being able to grow on trehalose and unable to use lactate and erythritol. Growth on mannitol, N-acetyl-glucosamine, potassium gluconate and sorbitol varied between strains.

Growth capacities on hydrophobic substrates

We assessed growth on lipids and alkanes by carrying out drop tests for 13 HS, initially on MMB (Table 2). Three of the 10 lipid sources tested supported the growth of all species (tributyrin, methyl myristate and methyl palmitate), two were not used by any of the species (hexanoic acid and methyl hexanoic, both C6 compounds) and the remaining five were used by some, but not all species. The strains differed in their ability to grow on the three alkanes (Figure 1). Thus, all strains were able to grow on at least some hydrophobic substrates. However, some displayed growth delay or an absence of growth on specific substrates. *YAAL* could not grow on C18:1 whatever the type of compound supplied (triglyceride, methyl or fatty acid) or on C22:1. On all other media, growth was delayed or weak, except in the presence of tributyrin. Similarly, *YAYA* did not grow on the three types of C18:1 or on C10 (methyl decanoate). On all other lipids except tributyrin, growth was delayed or weak and hexadecane (C16) was the only alkane of the three tested that supported the growth of this strain. Surprisingly, *YALI* formed no colonies on alkane minimal media under these growth conditions.

There are three possible reasons for the inability of a strain to grow on a given substrate: the substrate may not be taken up by the cells, it may not be metabolized or it may have a toxic effect, preventing growth. We tested the hypothesis of toxicity by repeating the drop tests on a rich medium (YP, yeast extract plus peptone) supplemented with hydrophobic compounds (Figure 2, Figure 3 and additional Figure S2). Under these conditions, all strains could grow in the presence of all the hydrophobic substrates tested, with two exceptions, as described below. The results of these tests indicated that some hydrophobic substrates neither supported nor inhibited cell growth. The two exceptions were the two C6 compounds (hexanoic acid and methyl hexanoic acid), which strongly inhibited all strains, although *YAPH* and *YAHI* were able to grow on YP-methyl hexanoic after 20 days. Similarly, on MMB-methyl decanoate, no growth defect was observed at low cell density, whereas growth was strongly inhibited at high cell density. No such growth inhibition was observed on YP-methyl decanoate (Figure 2).

Morphology on plates was very variable, ranging from smooth to rough colonies, and from superficial to agar-invasive forms, depending on the medium (see figures 2, 3 and data not shown). The presence of a halo around cells grown on plate containing insoluble substrates generally reflects the secretion of surfactants or of extracellular enzymes, such as lipases or esterases [39]. Only

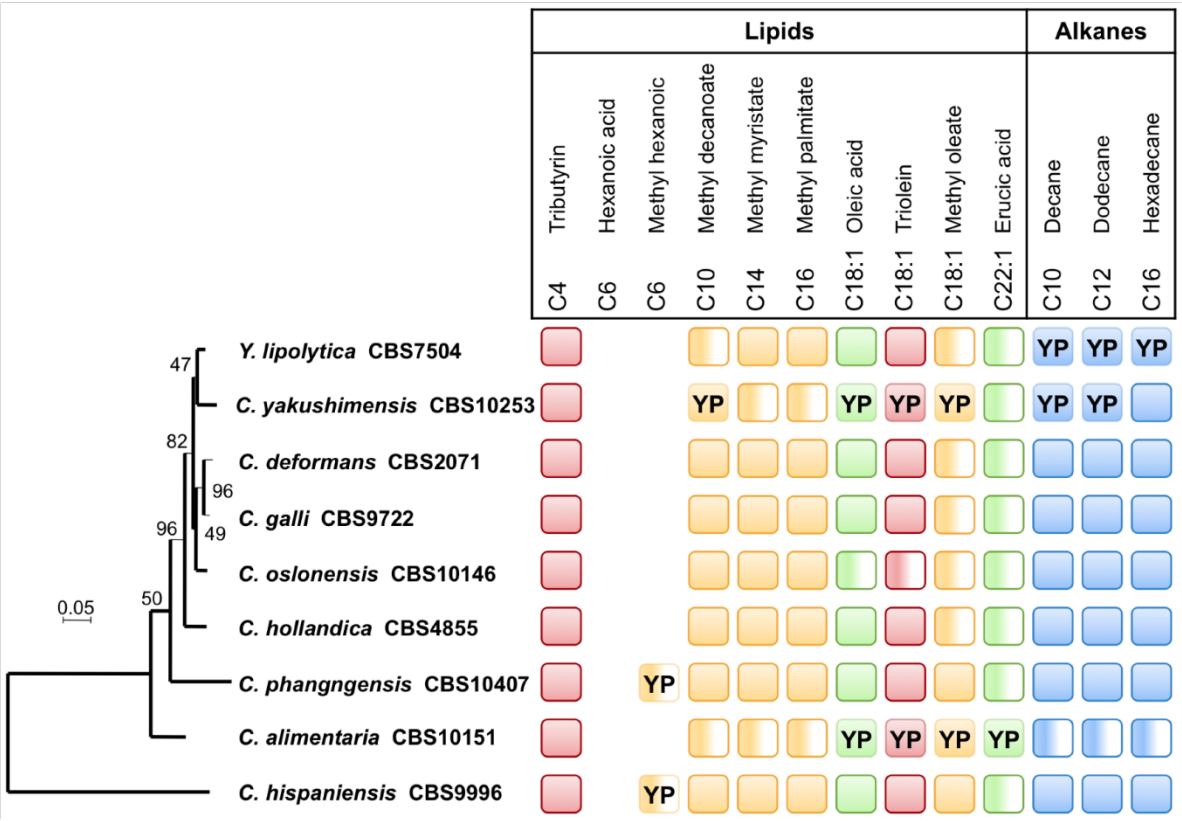


Figure 1. Growth capacities of *Yarrowia* strains on 13 hydrophobic substrates as a function of their phylogenetic position. Squares represent growth capacity on minimal and rich (YP) media supplemented with triglycerides in red, methylates in yellow, free fatty acids in green and alkanes in blue. The squares marked YP reflect the capacity to grow on the YP base used as a carbon source, in the presence of the corresponding HS. Half-colored squares indicate slow or weak growth. The absence of a square indicates an absence of growth on both YP and minimal media. The phylogenetic tree was constructed by the concatenation of seven markers (2204 amino acids) and its robustness was estimated with a bootstrap of 100 replicates.

doi:10.1371/journal.pone.0063356.g001

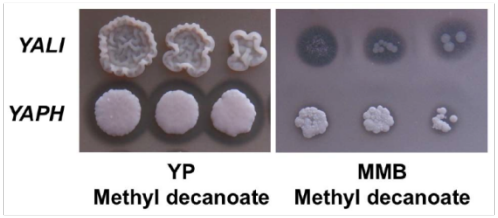


Figure 2. Drop tests on methyl decanoate in MMB and YP media, after 9 days of culture. Three spots are shown, corresponding to approximately 25 to 625 cells, from right to left. Strains are represented by an abbreviation of their species name, as shown in Table 1.
doi:10.1371/journal.pone.0063356.g002

four of the 13 HS tested on MMB and YP media were associated with halo formation: tributyrin, methyl decanoate, oleic acid and triolein (Additional Table S3). Some strains formed a halo on MMB or YP tributyrin plates, whereas others did not (Figure 3). The presence of a halo was not associated with colony size and depended on the medium, *i.e.* MMB or YP. For instance, *YAOS* formed a large halo on YP-tributyrin but grew less well than *YAHI*, which did not form a halo.

Lipid storage capacities are variable within the *Yarrowia* clade

The best way to determine the storage capacity of yeasts with minimal bias due to *de novo* synthesis is to estimate their lipid content on a lipid-based growth medium. We determined growth characteristics, such as lag phase, generation time and maximal OD_{600}, on the standard medium used for this purpose in *Y. lipolytica*: an oleic acid medium supplemented with small amounts of yeast extract to facilitate the transition from sugar to hydrophobic carbon sources [17]. (Figure S3A). All strains grew in this medium, without a lag phase (Table 3). Generation time and maximum cell growth differed between species, ranging from 2.06 h to 3.91 h and from 17.6 to 53.1 OD_{600}, respectively. *YAHI* was the fastest growing species on oleic acid, reaching 53.10 OD_{600} units after about 23 h of growth, with a generation time of 2.06 h. By contrast, *YAGA* seemed be assimilate oleic acid much less efficiently, with a generation time almost twice that of *YAHI*. We were therefore unable to use the standard protocols for estimating lipid storage in *Y. lipolytica*. Thus, to ensure that the data obtained were comparable, we had to identify similar physiological stages not necessarily corresponding to the same time in culture.

We estimated the maximal level of lipid storage, by determining cellular lipid content at various time points. Based on the growth curves obtained, we selected three time points between the beginning of the exponential phase and the deceleration phase.

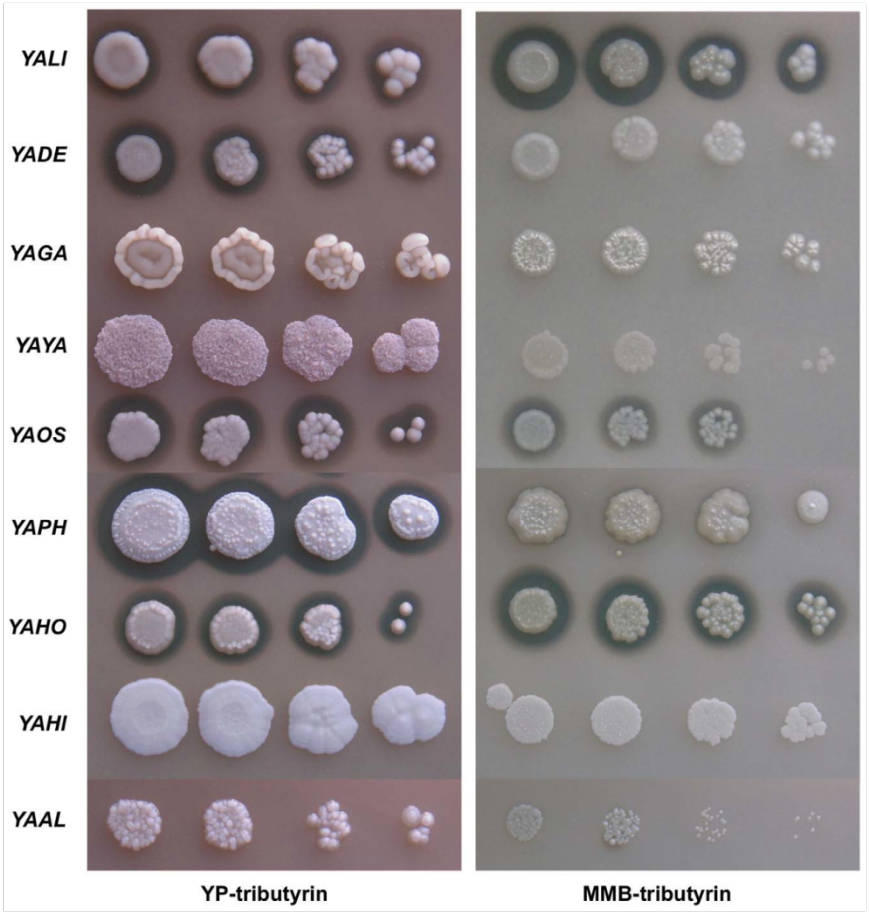


Figure 3. Drop tests on YP- and MMB-tributyrin media after 5 days of culture. Four spots are shown, with approximately 5 to 625 cells, from right to left. A halo is present around strains *YALI*, *YADE*, *YAOS*, *YAPH* and *YAHO*. Strains are represented by an abbreviation of their species name, as shown in Table 1.
doi:10.1371/journal.pone.0063356.g003

Table 3. Growth parameters on oleic acid and glucose media.

Strain	μmax (h^{-1})	Generation time (h)	Lag phase λ (h)	Maximum cell density (OD_{600})
Oleic acid				
YALI	0.21	3.27	-	35.31
YAYA	0.18	3.73	-	20.35
YADE	0.25	2.74	-	17.60
YAGA	0.18	3.91	-	23.52
YAOS	0.22	3.12	-	19.98
YAHO	0.26	2.67	-	33.66
YAPH	0.32	2.18	-	46.8
YAAL	0.28	2.46	-	28.27
YAHI	0.34	2.06	-	53.10
Glucose				
YALI	0.44	1.56	2.54	25.00
YAYA	0.71	0.97	5.19	20.86
YADE	0.78	0.89	5.61	17.80
YAGA	0.42	1.64	3.38	23.71
YAOS	0.63	1.09	5.71	19.00
YAHO	0.53	1.32	3.85	27.86
YAPH	0.57	1.21	1.26	29.14
YAAL	0.64	1.08	8.77	23.2
YAHI	0.53	1.32	7.93	17.3

doi:10.1371/journal.pone.0063356.t003

These time points corresponded roughly to the time required to obtain one fifth of the maximum OD_{600} (OD max), half the OD max and the OD max. An example is provided in Additional Figure S4. For lipid extraction, we collected cells for two biological replicates at about the three time points derived from the growth curve (Additional Table S4).

The total lipid content of cells grown on oleic acid is shown in Figure 4A. Three different types of behavior were observed. The first was characterized by an increase in lipid accumulation during the exponential growth phase followed by a decrease in lipid levels during the deceleration phase, possibly corresponding to lipid reconsumption. This pattern was observed in *YALI*, *YAGA*, *YAAL* and *YAHI*. The second type of pattern consisted of a continuous increase in lipid content and was observed in *YAOS* and *YAPH*. The three remaining strains, *YADE*, *YAYA* and *YAHO*, displayed decreases in lipid content over time. For instance, the lipid content of *Y. deformans* CBS2071 decreased from 34.4% to 15.2% of CDW.

The maximal level of lipid storage differed considerably between species, ranging from 29.4% of CDW for *YAOS* to 67.1% of CDW for *YAHI*. *YALI* was the second most efficient strain for lipid storage, which accounted for 52.2% of CDW.

By contrast, overall lipid composition remained stable over time for all species (Additional Table S5). Most of the lipid was in the form of C18:1(n-9), as expected given that the main carbon source in the growth medium was oleic acid. In *YALI*, *YAGA*, *YAHI* and *YAHO*, the amount of C18:1(n-9) varied from 23.1 to 38.3 % of CDW, suggesting that oleic acid uptake was more efficient in these species than in the other strains tested (Additional Fig S5). The second most abundant compound accumulating was generally C16:1(n-7), but its level was always low and varied from 0.7 to 11.8 % of CDW.

De novo lipid synthesis

We then determined the capacity for *de novo* lipid synthesis of the nine strains. This capacity was estimated on a glucose medium supplemented with yeast extract: a rich medium devoid of hydrophobic compounds. As in evaluations of storage capacity, we first established growth curves (Additional Figure S3B). By contrast to what was observed on oleic acid, there was a conspicuous lag phase for all strains, lasting 1.26 to 8.77 h, and generation time was clearly shorter than on oleic acid media, but remained variable (0.89 to 1.64 h). OD max values were similar on the two media for all strains except *YALI*, *YAPH* and *YAHI*, for which the OD max on glucose was lower than that on oleic acid, by a factor of up to three. For instance, the OD max for *YAHI* was 17.3 on glucose and 53.1 on oleic acid.

We then measured lipid contents at three time points defined as described above: 1/5, 1/2 and 1 OD max (Additional Table S4). Almost all the strains presented similar patterns of total lipid content, with large increases during the deceleration phase (Figure 4B). However, slight differences in kinetics were observed. For example, in *YAHI*, the increase in lipid content was more progressive, remaining significant during the deceleration phase (from 5.1 to 7.2% of CDW) whereas the increase was less pronounced for *YAAL* (from 4.0 to 4.5% of CDW). There were two exceptions to this rule: *YALI*, the lipid content of which remained stable over time, and *YAYA*, the lipid content of which decreased gradually from 6 to 4.7% of CDW.

In most strains, C18:1(n-9) was the major compound that accumulated, its levels increasing over the entire time course. However, in *YAAL*, *YAYA* and *YAHI*, this compound was not the principal compound at OD max. In *YAAL* and *YAYA*, C18:2(n-6) was the major compound at all three time points considered. In

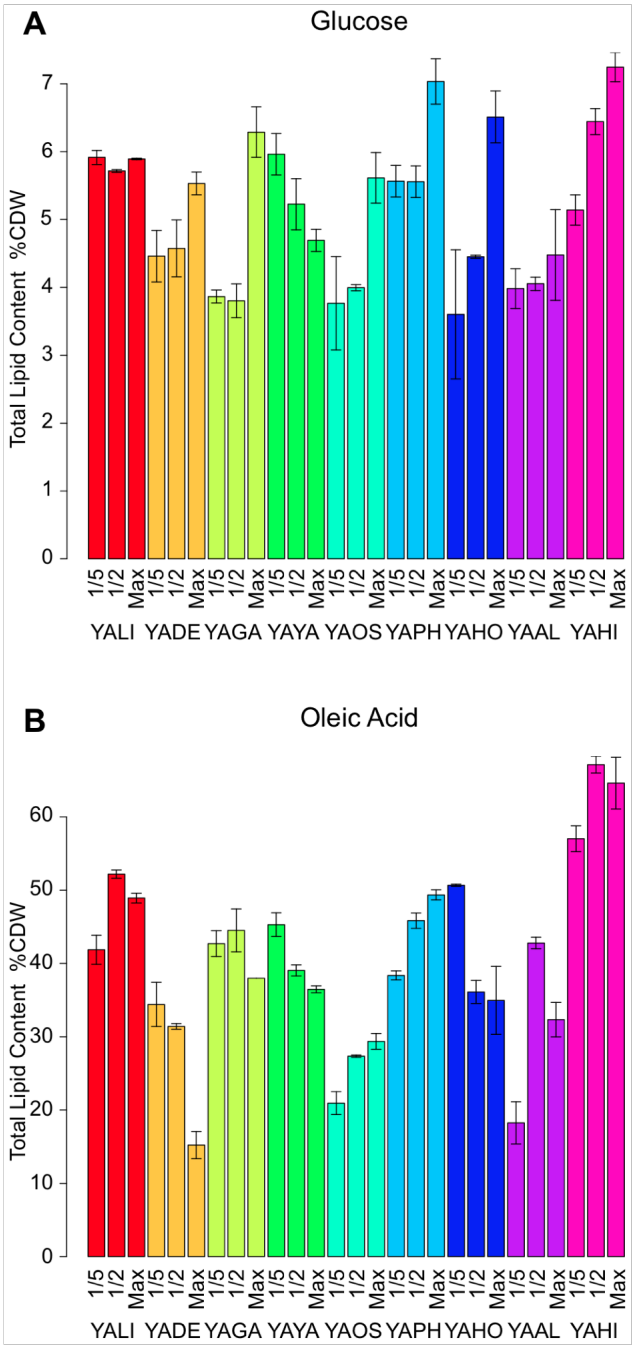


Figure 4. Lipid content (% CDW) of *Yarrowia* strains over time. For each species, the histogram represents three different physiological states: 1/5 of maximum growth, 1/2 maximum growth and maximum growth. Error bars indicate the deviation from the mean deduced from two replicates. Analyses were performed on both glucose (A) and oleic acid (B). Strains are represented by an abbreviation of their species name, as shown in Table 1.
doi:10.1371/journal.pone.0063356.g004

YAHI, the increase in lipid content was due to both C16:0 and C18:1(n-9) (Figure 5, Additional Fig S5, Additional Table S5).

Discussion

We analyzed eight closely related species of the *Yarrowia* clade, to obtain preliminary data concerning their physiology and, in particular, their ability to grow on and to assimilate HS, before beginning comparative genomic and transcriptomic studies. All the data discussed below were obtained from a single isolate per species and may, therefore, not be representative of the entire species.

As expected on the basis of published findings, despite their isolation from highly diverse environments (Table 1), these strains displayed a very limited range of non hydrophobic substrate used, with low levels of variability[18,19,20,21]. Glucose, fructose and glycerol were the only non hydrophobic carbon sources used by all strains of the clade. The screening of these strains for the detection of enzymatic activities useful for the exploitation of agricultural by-products, such as xylose degradation, does not therefore appear to be pertinent.

As *Y. lipolytica* uses HS efficiently, we tested the eight strains on 13 different substrates (alkanes, fatty acids and their methylated derivatives, and triglycerides) with carbon chains of different lengths (C4 to C22). Most of these substrates had never before been tested as substrates for species of this clade. We observed differences in growth on these substrates, the size of the hydrolysis halo and sensitivity to short-chain fatty acids between the species tested. These fatty acids were toxic to all strains, but *YAPH* and *YAHI* were able to grow on YP in the presence of C6 after a very long time lag.

Toxicity was clearly observed for *YALI* on MMB-methyldecanoate, but not on YP methyldecanoate (Figure 2). This may result from the balance between transport and degradation on the one hand and the hydrolysis of methyl-C10 by lipases/esterases to release decanoate, a compound toxic to cells, on the other. At high cell density, methyl-C10 hydrolysis by lipases may occur too rapidly, releasing too much toxic C10 for the cells to cope with, whereas, at lower cell density, the release of toxic C10 might be slow enough to prevent cell intoxication but permit cell growth. The halo was smaller for *YAPH* on the same medium, which thus released C10 at a rate compatible with growth. On YP supplemented with methyl-C10, both strains grew vigorously. All strains of the *Yarrowia* clade were able to grow efficiently on YP in the absence of any other carbon source (see additional Figure S2). On YP-methyl decanoate, no hydrolysis halo was observed with *YALI*, suggesting that the corresponding lipase/esterase is repressed in the presence of peptone. This putative repression was not observed with *YAPH*, suggesting that cell intoxication was prevented either by slow C10 uptake or by the rapid β-oxidation of C10. Further studies involving comparative genomics and transcriptomics would make it possible to test these hypotheses. These approaches might also shed light on the evolution of the *POX* and *LIP* gene networks in *YAPH* and *YALI*, in which six different acyl-CoA oxidases (Pox1p to Pox6p) of different chain length specificities have been described [40,41,42], together with 16 lipases and four esterases with different patterns of regulation and chain-length specificities [7].

We tried to determine whether these yeasts were oleaginous, by analyzing lipid accumulation and synthesis. However, the concept of what constitutes an oleaginous yeast remains unclear. Some authors consider that microorganisms should be able to accumulate more than 20% of CDW in the form of lipid to be considered oleaginous, whereas others consider that oleaginous organisms should be able to synthesize lipids from non HS carbon sources and to store them [8]. This is a crucial difference as some species are probably able to synthesize, but not to accumulate lipids efficiently, whereas other species may only accumulate lipids from the extracellular medium efficiently. These two capacities should thus be estimated independently, on different carbon sources, such as glucose for lipid synthesis and a model oleic acid medium

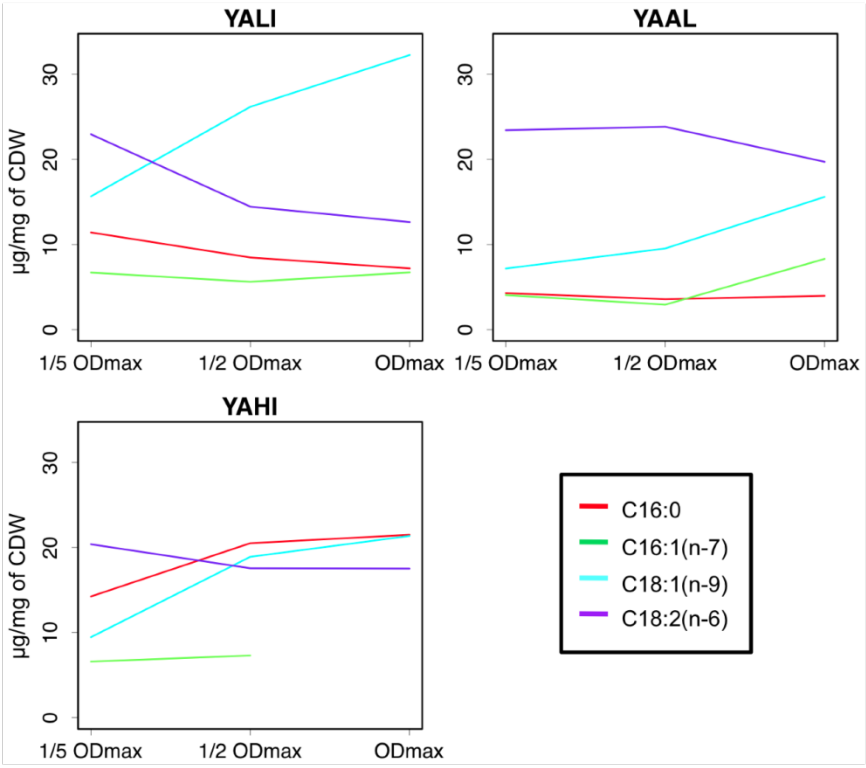


Figure 5. Changes in cellular lipid composition over time in strains grown on glucose media. *YALI*: *Yarrowia lipolytica* W29, *YAAL*: *C. alimentaria* CBS10151, *YAHI*: *C. hispaniensis* CBS9996. Quantities are in mg/g CDW.
doi:10.1371/journal.pone.0063356.g005

previously used for *YALI* for the assessment of storage capacities [16].

Our results demonstrated that all strains cultured on oleic acid accumulated more than 30% of their CDW in the form of lipid at least one time point. Our findings suggest that all species of the clade are oleaginous. Most accumulated smaller amounts of lipid than *YALI*, but at least one, *YAHI*, accumulated almost 30% more lipid than *YALI*. However, different accumulation profiles were observed during the growth phase: some strains progressively accumulated triglycerides (*YAOS* and *YAPH*), others remobilized lipids (*YADE*, *YAGA* and *YAHO*), and lipid content remained essentially stable in *YALI* and *YAAL*. Lower and less variable levels of lipid accumulation were observed on glucose. Depending on the strain, the major compound accumulated was C18:1, C18:2 or C16:0. No molecules with a longer chain length were observed, by contrast to what has been reported for *Saccharomyces cerevisiae* and *Hansenula polymorpha* [43,44]. *Candida albicans* can synthesize and accumulate C18:3 to levels of up to 22.7% of total fatty acids [43], but none of the yeasts of the *Yarrowia* clade was able to synthesize C18:3. This finding is consistent with the absence from the nuclear genomes of *YAGA*, *YAYA*, *YAAL*, *YAPH* and *YAHI* of a gene encoding a Δ15-desaturase, whereas homologs of YALI0C05951g (Δ9-desaturase) and YALI0B10153g (Δ12-desaturase) were identified (unpublished data). The variable levels of the various compounds may reflect differences in the activity or expression of the Δ12-desaturase (for *YAAL*, *YAYA*) or elongases (for *YAHI*).

The major conclusion of this work is that all species of the *Yarrowia* clade are oleaginous, but that they differ in their profiles of HS use and lipid accumulation. These data should facilitate the development of more robust, better performing strains for lipid accumulation and for fatty acid profile modification. Our findings also suggest that there are metabolic and genetic differences between the strains of the clade studied here, possibly reflecting species-specific differences in gene content and/or regulation. *C. hispaniensis*, which grew well on diverse substrates and displayed a high storage capacity, appears to be a promising model for biotechnological applications *per se* or as a source of genes for the improvement of *Y. lipolytica*. Further comparative genomic and transcriptomic studies are underway in the *Yarrowia* clade, to establish correlations with phenotypic variability and to obtain evidence for their genetic bases.

Supporting Information

Figure S1 Growth curves of YAAL on YPD at different temperatures. The growth curves at 15°C, 21°C and 25°C are shown as blue squares, green dots, and red triangles, respectively.
(PDF)

Figure S2 Drop tests on YP (A) and MMB (B) media with alkanes and methyl-esters of various chain lengths. Only one spot is presented for strains *YALI*, *YADE*, *YAGA*, *YAYA* and *YAOS*. Strains are represented by an abbreviation of their species name, as shown in Table 1.
(PDF)

Figure S3 Growth curves established on 2% oleic acid (A) and 2% glucose (B) media for the strains of the *Yarrowia* clade, over a period of 50 h. Strain names are abbreviated as follows: *YALI* (*Y. lipolytica* W29), *YAYA* (*Y. yakushimensis* CBS10253), *YADE* (*Y. deformans* CBS2071), *YAGA* (*C. galli* CBS9722), *YAOS* (*C. oslonensis* CBS10146), *YAHO* (*C. hollandica* CBS4855), *YAPH* (*C. phangngensis* CBS10407), *YAAL* (*C. alimentaria* CBS10151), *YAHI* (*C. hispaniensis* CBS9996).
(PDF)

Figure S4 Growth curve and time points used for *YAGA* cultured on glucose (2%). The experimental growth curve is shown in red. Horizontal dotted lines correspond to the maximum OD, 1/2 OD max and 1/5 of OD max. Circles, triangles and diamonds indicate the OD at the time points used for lipid accumulation tests, at 4.6 h, 11.6 h and 23.7 h of culture, respectively.
(PDF)

Figure S5 Changes in cellular lipid composition over time for strains grown on oleic acid (A) and glucose (B). Quantities are in mg/g CDW. For each species, three different physiological states are represented: 1/5 the OD max (1/5), 1/2 the OD max (1/2) and the OD max (max).
(PDF)

Table S1 Growth capacities of the species of the *Yarrowia* clade on non hydrophobic carbon.
(XLSX)

Table S2 Primers used to amplify the protein-coding genes.
(XLSX)

Table S3 Presence of a halo on solid hydrophobic media.
(XLSX)

Table S4 Characteristics of growth over time for cultures on oleic acid and glucose media: time points, optical density, lipid content and cell dry weight.
(XLSX)

Table S5 Profile of fatty acid accumulation (% of CDW) on glucose and oleic acid media.
(XLS)

Acknowledgments

We thank Thanos Beopoulos, Vincent Sauveplane and Thierry Dulermo for technical assistance with lipid content analysis and stimulating discussions.

Author Contributions

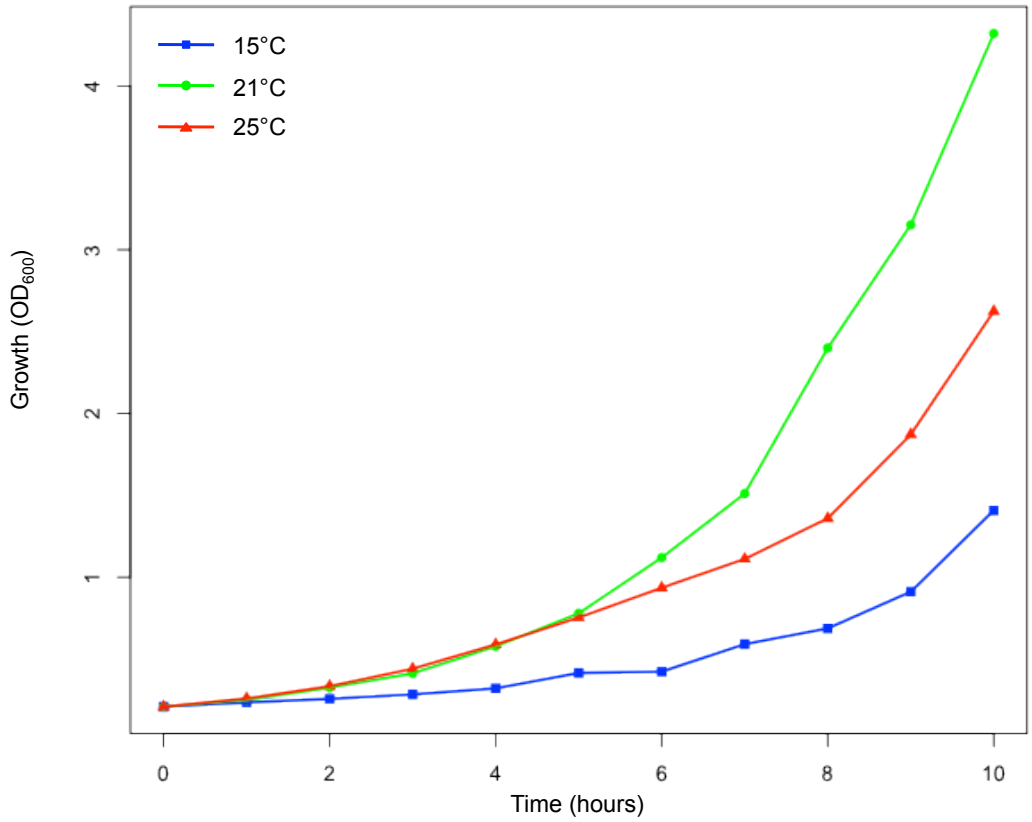
Conceived and designed the experiments: SM CN. Performed the experiments: SM CN. Analyzed the data: SM CG JMN CN. Wrote the paper: SM CG JMN CN.

References

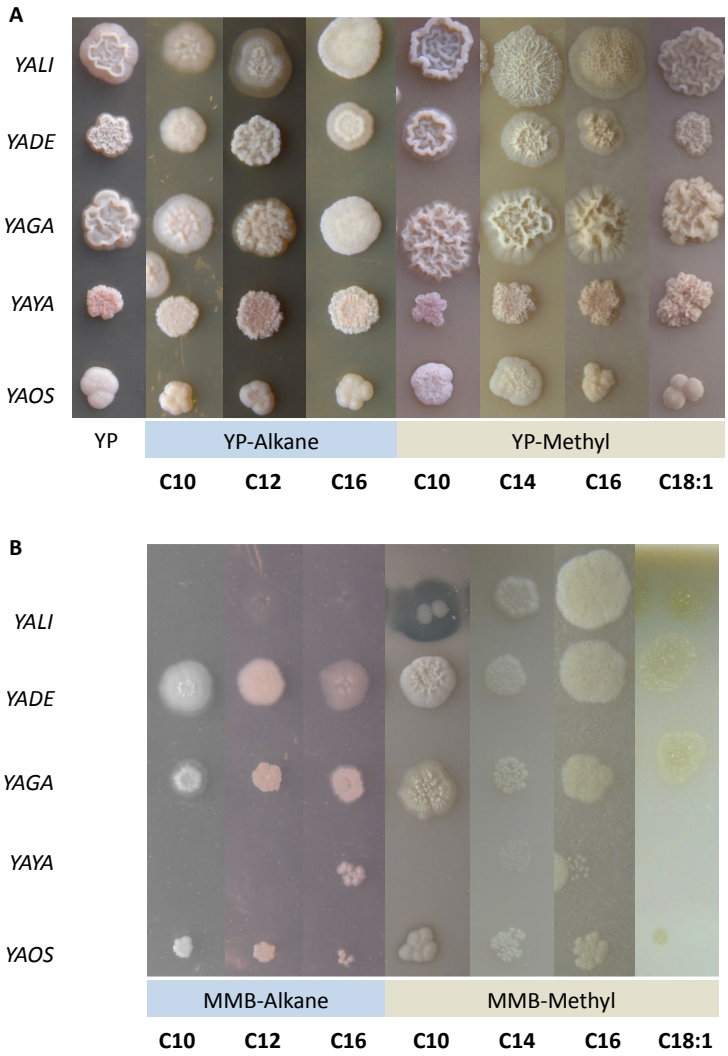
1. Ratledge C (1991) Microorganisms for lipids. Acta Biotechnologica 11: 429–438.
2. Poncet S, Arpin M (1965) Candida species without fermentative capacity (Cryptococcoceae). Antonie Van Leeuwenhoek 31: 433–464.
3. Barnett JA, Payne RW, Yarrow D (2000) Yeasts: Characteristics and Identification. Cambridge: Cambridge University Press. 1150 p.
4. Klug MJ, Markovetz AJ (1967) Degradation of Hydrocarbons by Members of the Genus Candida II. Oxidation of n-Alkanes and 1-Alkenes by *Candida lipolytica*. Journal of Bacteriology 93: 1847–1852.
5. Nyns EJ, Chiang N, Wiaux AL (1968) Comparative lipid content of *Candida lipolytica* grown on glucose and onn-hexadecane. Antonie Van Leeuwenhoek 34: 197–204.
6. Nicaud JM (2012) *Yarrowia lipolytica*. Yeast 29: 409–418.
7. Fickers P, Marty A, Nicaud JM (2011) The lipases from *Yarrowia lipolytica*: genetics, production, regulation, biochemical characterization and biotechnological applications. Biotechnol Adv 29: 632–644.
8. Beopoulos A, Chardot T, Nicaud JM (2009) *Yarrowia lipolytica*: A model and a tool to understand the mechanisms implicated in lipid accumulation. Biochimie 91: 692–696.
9. Fillaudeau L, Cescut J, Anne-Archard D, Nicaud JM, Uribelarrea JL, et al. (2009) Morphology and rheological behaviour of *Yarrowia lipolytica* during production of intra-cellular energetic molecules: impact of lipid accumulation and genetic modifications. 8th World Congress of Chemical Engineering, WCCE8, Montréal, Canada.
10. Papanikolaou S, Aggelis G (2002) Lipid production by *Yarrowia lipolytica* growing on industrial glycerol in a single-stage continuous culture. Bioresource Technology 82: 43–49.
11. Papanikolaou S, Muniglia L, Chevalot I, Aggelis G, Marc I (2002) Yarrowia lipolytica as a potential producer of citric acid from raw glycerol. J Appl Microbiol 92: 737–744.
12. Papanikolaou S, Chevalot I, Komaitis M, Marc I, Aggelis G (2002) Single cell oil production by Yarrowia lipolytica growing on an industrial derivative of animal fat in batch cultures. Appl Microbiol Biotechnol 58: 308–312.
13. Papanikolaou S, Chatzifragkou A, Fakas S, Galiotou-Panayotou M, Komaitis M, et al. (2009) Biosynthesis of lipids and organic acids by *Yarrowia lipolytica* strains cultivated on glucose. European Journal of Lipid Science and Technology 111: 1221–1232.
14. Chi Z, Zheng Y, Jiang A, Chen S (2011) Lipid Production by Culturing Oleaginous Yeast and Algae with Food Waste and Municipal Wastewater in an Integrated Process. Applied Biochemistry and Biotechnology 165: 442–453.
15. Hong S-P, Seip J, Walters-Pollak D, Rupert R, Jackson R, et al. (2012) Engineering *Yarrowia lipolytica* to express secretory invertase with strong FBA1IN promoter. Yeast 29: 59–72.
16. Beopoulos A, Mrozova Z, Thevenieau F, Le Dall MT, Hapala I, et al. (2008) Control of lipid accumulation in the yeast *Yarrowia lipolytica*. Appl Environ Microbiol 74: 7779–7789.
17. Mlíčková K, Roux E, Athenstaedt K, d'Andrea S, Daum G, et al. (2004) Lipid accumulation, lipid body formation, and acyl coenzyme A oxidases of the yeast *Yarrowia lipolytica*. Appl Environ Microbiol 70: 3918–3924.
18. Péter G, Dlauchy D, Vasdinyei R, Tornai-Lehoczki J, Deák T (2004) *Candida galli* sp. nov., a new yeast from poultry Antonie van Leeuwenhoek 86: 105-110-110.
19. Kurtzman CP (2005) New species and a new combination in the Hyphopichia and *Yarrowia* yeast clades. Antonie van Leeuwenhoek 88: 121–130.
20. Knutsen AK, Robert V, Poot GA, Epping W, Figge M, et al. (2007) Polyphasic re-examination of *Yarrowia lipolytica* strains and the description of three novel Candida species: *Candida oslonensis* sp. nov., *Candida alimentaria* sp. nov. and *Candida hollandica* sp. nov. Int J Syst Evol Microbiol 57: 2426–2435.
21. Limtong S, Youngmanitchai W, Kawasaki H, Seki T (2008) *Candida phangngensis* sp. nov., an anamorphic yeast species in the *Yarrowia* clade, isolated from water in mangrove forests in Phang-Nga Province, Thailand. Int J Syst Evol Microbiol 58: 515–519.
22. Kurtzman CP (2011) *Yarrowia* van der Walt & von Arx (1980). In: Kurtzman CP, Fell JW, Boekhout T, editors. The Yeasts, a taxonomic study. London: Elsevier. pp. 927–929.
23. Groenewald M, Smith M (2013) The teleomorph state of *Candida deformans* Langeron & Guerra and description of *Yarrowia yakushimensis* comb. nov. Antonie van Leeuwenhoek: 1–6.
24. Lachance M-A, Kurtzman CP (2011) *Lachancea* Kurtzman (2003). In: Kurtzman CP, Fell JW, Boekhout T, editors. The Yeasts, a taxonomic study. London: Elsevier. pp. 511–519.
25. Barth G, Gaillardin C (1996) *Yarrowia lipolytica*. In: Wolf WK, editor. Non-Conventional Yeasts in Biotechnology. Berlin: Springer-Verlag. pp. 313–388.
26. Bigey F, Tuery K, Bougard D, Nicaud J-M, Moulin G (2003) Identification of a triacylglycerol lipase gene family in *Candida deformans*: molecular cloning and functional expression. Yeast 20: 233–248.
27. Boutur O, Dubreucq E, Galzy P (1995) Factors influencing ester synthesis catalysed in aqueous media by the lipase from *Candida deformans* (zach) langeron and guerra. Journal of Biotechnology 42: 23–33.
28. Muderhwa JM, Ratomahenina R, Pina M, Graille J, Galzy P (1985) Purification and properties of the lipase from *Candida deformans* (zach) langeron and guerra. Journal of the American Oil Chemists Society 62: 1031–1036.
29. Kurtzman CP, Fell JW, Boekhout T, Robert V (2011) Methods for isolation, phenotypic characterisation and maintenance of yeasts. In: Kurtzman CP, Fell JW, Boekhout T, editors. The Yeasts, a taxonomic study. London: Elsevier. pp. 87–110.
30. Mauersberger S (1991) Mutants of alkane oxidation in the yeasts *Yarrowia lipolytica* and *Candida maltosa*. In: Sharyshev AA, Finogenova TV, editors. Alkane metabolism and oversynthesis of metabolites by microorganisms. Pushchino, USSR: Center for Biological Research, USSR Academy of Sciences. pp. 59–78.
31. R Development Core Team (2011) R: A language and environment for statistical computing. In: Computing RFfS, editor. Vienna, Austria.
32. Browse J, McCourt PJ, Somerville CR (1986) Fatty acid composition of leaf lipids determined after combined digestion and fatty acid methyl ester formation from fresh tissue. Analytical Biochemistry 152: 141–145.
33. Aguileta G, Marthey S, Chiapello H, Lebrun MH, Rodolphe F, et al. (2008) Assessing the performance of single-copy genes for recovering robust phylogenies. Syst Biol 57: 613–627.
34. Hoffman CS, Winston F (1987) A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of *Escherichia coli*. Gene 57: 267–272.
35. Bonfield JK, Smith K, Staden R (1995) A new DNA sequence assembly program. Nucleic Acids Res 23: 4992–4999.

36. Edgar R (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.
37. Thompson JD, Gibson TJ, Plewniak Fdr, Jeanmougin Fo, Higgins DG (1997) The CLUSTAL_X Windows Interface: Flexible Strategies for Multiple Sequence Alignment Aided by Quality Analysis Tools. Nucleic Acids Research 25: 4876–4882.
38. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696–704.
39. Pignede G, Wang H, Fudalej F, Gaillardin C, Seman M, et al. (2000) Characterization of an Extracellular Lipase Encoded by LIP2 in *Yarrowia lipolytica*. J Bacteriol 182: 2802–2810.
40. Wang H, Le Dall M-T, Waché Y, Laroche C, Belin J-M, et al. (1999) Cloning, sequencing, and characterization of five genes coding for Acyl-CoA oxidase isozymes in the yeast *Yarrowia lipolytica*. Cell Biochemistry and Biophysics 31: 165–174.
41. Luo YS, Nicaud JM, Van Veldhoven PP, Chardot T (2002) The acyl-CoA oxidases from the yeast Yarrowia lipolytica: characterization of Aox2p. Arch Biochem Biophys 407: 32–38.
42. Luo YS, Wang HJ, Gopalan KV, Srivastava DK, Nicaud JM, et al. (2000) Purification and characterization of the recombinant form of Acyl CoA oxidase 3 from the yeast *Yarrowia lipolytica*. Arch Biochem Biophys 384: 1–8.
43. Tylicki A, Siemieniuk M, Dobrzyn P, Ziolkowska G, Nowik M, et al. (2012) Fatty acid profile and influence of oxythiamine on fatty acid content in *Malassezia pachydermatis*, *Candida albicans* and *Saccharomyces cerevisiae*. Mycoses 55: e106–e113.
44. Prasitchoke P, Kaneko Y, Bamba T, Fukusaki E, Kobayashi A, et al. (2007) Identification and characterization of a very long-chain fatty acid elongase gene in the methylotrophic yeast, *Hansenula polymorpha*. Gene 391: 16–25.

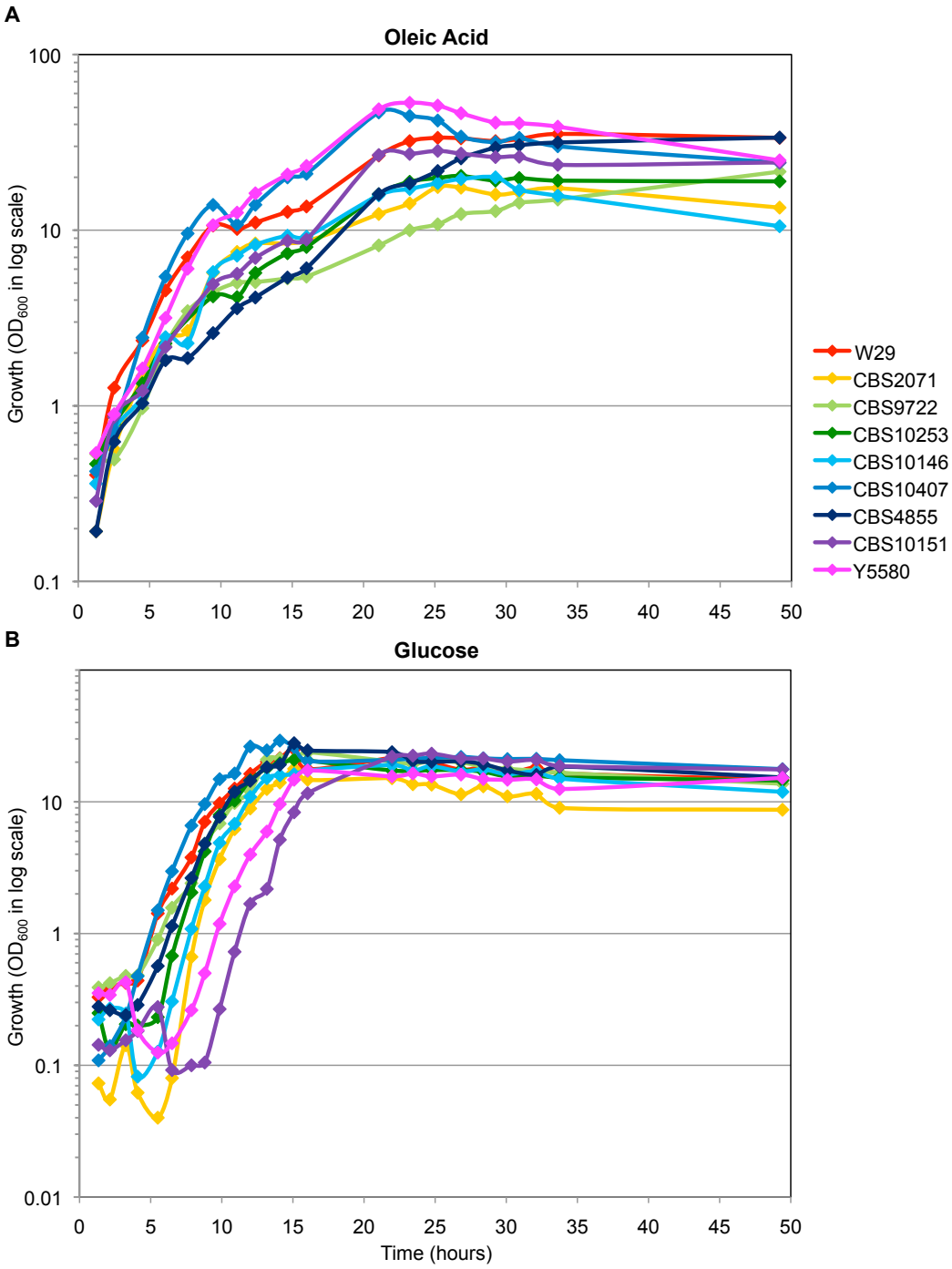
Supplementary data :



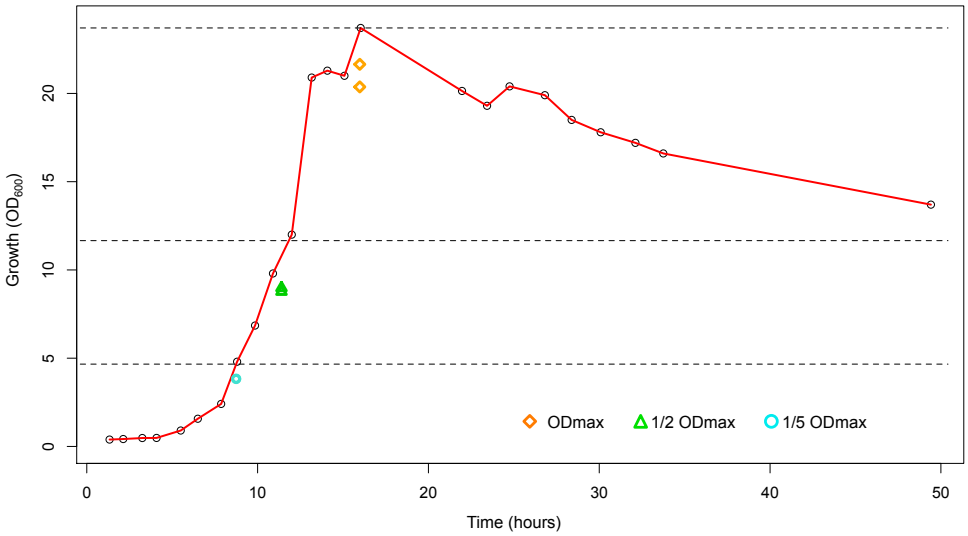
Additional Figure S1: Growth curves of YAAL on YPD at different temperatures. The growth curves at 15°C, 21°C and 25°C are represented in blue with squares, in green with dots, and in red with triangles, respectively.



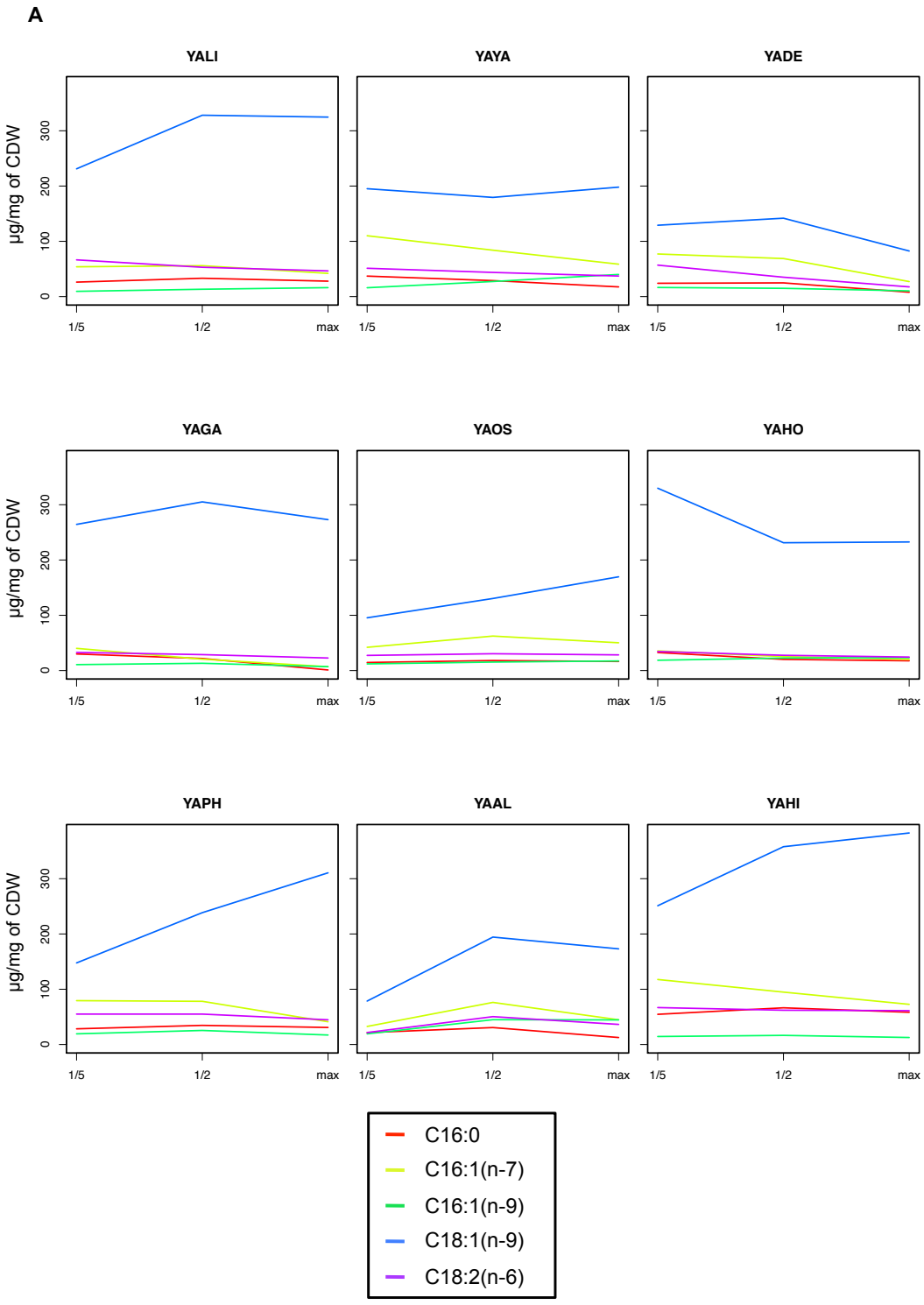
Additional Figure S2: Drop tests on YP (A) and MMB (B) media with alkanes and methyl-esters of different chain length. Only one spot is presented for strains *YALI*, *YADE*, *YAGA*, *YAYA* and *YAOS*. Strains are represented by an abbreviation of their species name as shown in Table 1.

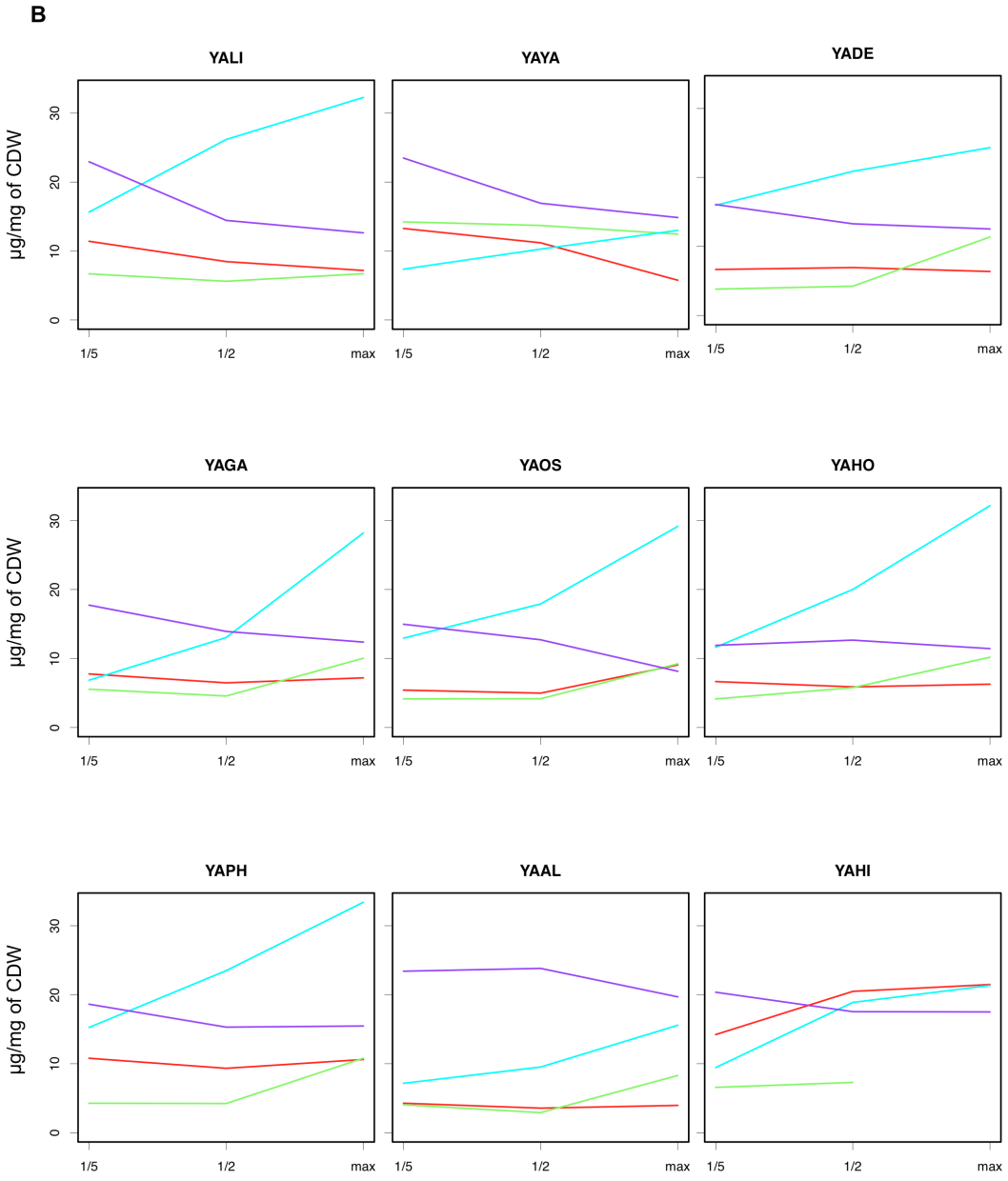


Additional Figure S3: Growth curves established on oleic acid 2% (A) and glucose 2% (B) media for the strains of the *Yarrowia* clade during 50 h. Strain names are abbreviated as follows: *YALI* (*Y. lipolytica* W29), *YAYA* (*C. yakushimensis* CBS10253), *YADE* (*C. deformans* CBS2071), *YAGA* (*C. galli* CBS9722), *YAOS* (*C. oslonensis* CBS10146), *YAHO* (*C. hollandica* CBS4855), *YAPH* (*C. phangngensis* CBS10407), *YAAL* (*C. alimentaria* CBS10151), *YAHI* (*C. hispaniensis* CBS9996).



Additional Figure S4: Growth curve and time points used for *YAGA* cultivated on Glucose (2%). Experimental growth curve is in red. Horizontal dotted lines correspond to the maximum OD, to the 1/2 and to the 1/5 of OD max. Circles, triangles and diamonds indicate OD at the time points used for lipid accumulation tests at 4.6 h, 11.6 h and 23.7 h of culture, respectively.


A
YALI
YAYA
YADE
YAGA
YAOS
YAHO
YAPH
YAAL
YAHI
µg/mg of CDW
0
100
200
300
1/5
1/2
max
C16:0
C16:1(n-7)
C16:1(n-9)
C18:1(n-9)
C18:2(n-6)



Additional Figure S5: Evolution of cellular lipid composition over time of strains grown on oleic acid (A) and glucose media (B). Quantities are in mg/g CDW. For each species, three different physiological states are represented: 1/5 of the log phase (1/5), 1/2 of the log phase (1/2) and maximum growth (max).

Additional Table S1: Growth capacities of the species of the *Yarrowia* clade on non hydrophobic carbon compounds

	Y. lipolytica **CBS7504 - W29**			***C. yakushimensis*** **CBS10253**		***C. deformans*** **CBS2071**			***C. galli*** **CBS9722**			***C. oslonensis*** **CBS10146**		***C. phangngensis*** **CBS10407**			***C. hollandica*** **CBS4855**			***C. alimentaria*** **CBS10151**		***C. hispaniensis*** **CBS9996**	
	API	CBS	Biotek	API	Biotek	API	CBS	Biotek	API	CBS	Biotek	API	Biotek	API	CBS	Biotek	API	CBS	Biotek	API	Biotek	API	Biotek
Levulinate	-			-		-			-	-		-		-			-			-		-	
Arabinose	-	+	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
Cellobiose	-	-		-		-	-		-	-		-		-	-		-	-		-		-	
Esculin ferric citrate	-			-		-			-			-		-			-			-		-	
Galactose	-	-		-		-	-		-	w	-	-		-			-	-		-		-	
Glucosamine	-	-		-		d	-	-	-	-		-		-			-	-		-		-	
Inositol	-	-		-		-	-		-	-		-		-	-		-	-		-		-	
Lactose	-	-		-		-	-		-	-		-		-	-		-	-		-		-	
Maltose	-	-		-		-	-		-	w	-	-		-	-		-	-		-		-	
Melezitose	-	-		-		-	-		-	-		-		-	-		-	-		-		-	
Melibiose	-	-		-		-	-		-	-		-		-	-		-	-		-		-	
Methyl α-D-glucopyranoside	-			-		-			-			-		-			-			-		-	
Palatinose	-			-		-			-	w		-		-			-			-		-	
Potassium-2-ketogluconate	-			-		-			-			-		-			-			-		-	
Raffinose	-	-		-		-	-		-	-		-		-	-		-	-		-		-	
Rhamnose	-	-		-		-	-		-	-		-		-	-		-	-		-		-	
Ribose	-	+	-	-	-	d,w	d	-	-	-	-	-	-	-		-	-	-,d,w	-	d	-	-	-
Sucrose	-		-	-	-	-		-	-		-	-	-	-		-			-	-	-	-	-
Sodium glucuronate	-			-		-			-			-		-			-			-		-	
Sorbose	-	-		-		-	-		-	+	-	-		-	+	-	-	d	-	-		-	
Xylose	-	-	-	-	-	+	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-
Cycloheximide	+	+		+		+	+		+	+		+		+	+		+	+		+		+	
Glucose	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Glycerol	+	+		+		+	+		+	-	+	+		+	+		+	+		+		+	
Fructose			+		+			+			+		+			+			+		+		+
Mannitol	+	+		d		d	+	-	+	-	-	+		+	+		+	+		d		+	
N-Acetyl-Glucosamine	+	-	+	+		+	-	+	-	-		+		-			+		+	+		-	
Potassium gluconate	+			+		+			-			d,w		-			-			d		-	
Sorbitol	-			d		d,w			d			+		-			-			+		d,w	
Lactate	+	+		+		d,w	d,w,+		+	-	-	+		+	+		+	+		d		-	
Erythritol	+	+		+		+	+		+	+		+		+	+		+	+		+		-	
Trehalose	-	-		-		-	-		-	-		-		-	-		-	-		-		+	

Additional Table S2: Primers used for amplification of the protein-coding genes.

YALI gene	Primer name	Sequence
YALI0B18722g	MS456_F	TTACACYGTGACAAGTGTG
	MS456_R	ATACCAGCCTTGGARATGGA
YALI0B08756g	MS277_2F	AKGGTGTCAACTGGCGWGAT
	MS277_2R	CACCGCCTTGAACCAGTTG
YALI0A00264g	MS444_F	CAACATYAAGGAHGGCCGATT
	MS444_R	CRATGTCCTTCTTGAGCTTCT
YALI0B16434g	MS561_F	GATAACGAGGCAATTGCCAT
	MS561_R	GATTCGCACCATCCACAGG
YALI0A02695g	FG598_F	GTGTGGTACKSTCACAACGG
	FG598_R	ACATCCTTGGCWTCCTGSCC
YALI0D11220g	FG610_1F	TTCAAGCAGGGATAYCAGAC
	FG610_R	ACACCRTCATCAATGGCYCGC
YALI0C09141g	EF1U1	CCACTGGTCACCTTATCTAC
	EF1F1	AACGGTCTGTCGCATGTCTC

Degenerate primers were used to allow PCR amplification of the *YALI* homologues in each of the eight species of the *Yarrowia* clade

Additional Table S3: Presence of halo on solid hydrophobic media

Strains[a]	tributyrin (C4)		methyl decanoate (C10)		oleic acid (C18:1)		triolein (C18:1)	
	YN	YP	YN	YP	YN	YP	YN	YP
YALI	++	+	+++	-	-	++	-	-
YAYA	-	-	-	-	-	-	-	-
YADE	+/-	++	-	-	-	-	-	-
YAGA	-	-	-	-	-	++	-	-
YAOS	++	++	-	-	-	-	-	-
YAHO	++	+	-	-	-	+	-	-
YAPH	+	+++	+/-	++	-	++	-	+
YAAL	+	-	-	-	-	-	-	+
YAHI	-	-	+/-	++	-	-	-	-

a) Strains are indicated as: *YALI* = *Y. lipolytica* W29, *YAKU* = *C. yakushimensis* CBS10253, *YADE* = *C. deformans* CBS2071, *YAGA* = *C. galli* CBS9722, *YAOS* = *C. oslonensis* CBS10146, *YAHO* = *C. hollandica* CBS4855, *YAPH* = *C. phangngensis* CBS10407, *YAAL* = *C. alimentaria* CBS10151, *YAHI* = *C. hispaniensis* CBS9996

Additional Table S4: Characteristic of growth cultures over time in oleic acid and glucose media: tir content and cell dry weight.

Strains	Physiolo-gical stages	Oleic acid				
		Time (hour)	Mean OD	Mean lipid content (% CDW)	Cell dry weight* (mg)	Time (hour)
YALI	1/5	8.2	8.4 (±0.1)	41.9 (±2.0)	5.3 (±0.4)	8.1
	1/2	17.3	30.4	52.2 (±0.6)	4.5 (±0.1)	10.7
	Max	25.3	34.3 (±1.9)	48.9 (±0.7)	5.0 (±0.2)	15.0
YADE	1/5	7.8	8.1 (±1.9)	34.4 (±3.0)	5.3 (±0.2)	9.8
	1/2	14.6	19.1 (±0.4)	31.4 (±0.4)	5.1 (±0.1)	11.7
	Max	25.1	29.5 (±0.1)	15.2 (±1.9)	4.4	15.0
YAGA	1/5	10.2	5.7 (±0.4)	42.7 (±1.8)	6.3 (±0.1)	8.7
	1/2	25.3	9.7 (±2.2)	44.5 (±2.9)	11.6 (±2.0)	11.4
	Max	36	18.9 (±0.5)	37.9	9.0 (±0.4)	16.0
YAYA	1/5	9.7	3.6 (±0.3)	45.3 (±1.6)	6.7 (±0.1)	8.8
	1/2	17.1	15.6 (±0.6)	39.0 (±0.7)	5.1 (±0.1)	10.7
	Max	26.4	24.5 (±0.8)	36.5 (±0.5)	3.5 (±0.3)	15.0
YAOS	1/5	8.6	2.4 (±0.2)	21.0 (±1.6)	5.5 (±0.4)	9.5
	1/2	16.3	14.9 (±0.1)	27.4 (±0.2)	5.2 (±0.2)	11.5
	Max	28.7	15.2 (±1.3)	29.4 (±1.1)	4.4 (±0.3)	16.0
YAPH	1/5	7.5	11.4 (±0.5)	38.4 (±0.6)	4.8 (±0.1)	7.5
	1/2	15.5	26.9 (±2.3)	45.8 (±1.0)	3.6 (±0.3)	9.5
	Max	21.1	21.4 (±0.5)	49.3 (±0.7)	7.1	14.1
YAHO	1/5	15.9	12.2 (±0.3)	50.7 (±0.1)	9.4 (±0.5)	9.1
	1/2	21.8	30.2 (±1.9)	36.1 (±1.6)	5.1 (±0.5)	11.5
	Max	33.5	22.8 (±0.1)	35.0 (±4.6)	6.2 (±0.3)	15.1
YAAL	1/5	11.0	2.8 (±0.5)	18.2 (±2.9)	9.9 (±0.6)	14.0
	1/2	17.0	18.5 (±0.8)	42.8 (±0.8)	4.6	15.8
	Max	25.7	23.2 (±1.0)	32.3 (±2.3)	4.7 (±0.3)	24.7
YAHI	1/5	9.5	15.4 (±0.5)	57.0 (±1.8)	3.4	11.7
	1/2	15.8	33.7 (±0.9)	67.1 (±1.2)	3.5 (±0.2)	13.8
	Max	23.2	41.6 (±3.7)	64.6 (±3.5)	4.4 (±0.3)	16.0

* Cell dry weight is given for an equivalent of 10 $0D_{600}$.
The absolute deviation to the mean is given within brackets when different from ±0.0

Additional Table S5: Profile of fatty acid accumulation (% of CDW) on glucose and oleic acid media.

Strain	Stage	C16:0	C16:1(n-7)	C16:1(n-9)	C18:1(n-9)	C18:1(n-9)t	C18:2(n-6)
Oleic acid							
YALI	1/5	2.6 (±0.1)	5.4 (±0.2)	1.0	23.1 (±1.1)	1.0	6.6 (±0.1)
	1/2	3.3	5.6	1.3	32.8 (±0.5)	1.6	5.3
	max	2.8 (±0.2)	4.2 (±0.1)	1.6	32.5 (±0.4)	1.4	4.6
YADE	1/5	2.4 (±0.1)	7.7 (±0.4)	1.7 (±0.2)	12.9 (±1.4)	0.5	5.7 (±0.6)
	1/2	2.5	6.9	1.5	14.2 (±0.2)	0.5 (±0.1)	3.5
	max	0.8 (±0.2)	2.7 (±0.3)	1.0 (±0.1)	8.3 (±1.0)	0.3	1.8 (±0.2)
YAGA	1/5	3.0	4.0	1.1	26.4 (±1.2)	1.3 (±0.1)	3.3 (±0.1)
	1/2	2.2	2.1	1.3 (±0.2)	30.5 (±2.3)	1.8 (±0.2)	2.9 (±0.3)
	max	0.1	0.7	0.7	27.3 (±0.1)	0.9	2.3 (±0.1)
YAYA	1/5	3.7 (±0.1)	11 (±0.4)	1.6 (±0.1)	19.5 (±0.9)	0.8 (±0.1)	5.1 (±0.3)
	1/2	2.9	8.4	2.7 (±0.1)	17.9 (±0.5)	0.5	4.4 (±0.1)
	max	1.8	5.9 (±0.3)	4.0	19.8 (±0.1)	0.6	3.7
YAOS	1/5	1.5 (±0.1)	4.2 (±0.2)	1.2	9.5 (±1.3)	0.4 (±0.2)	2.7
	1/2	1.8 (±0.1)	6.2 (±0.3)	1.5 (±0.1)	13.0 (±0.7)	0.5	3.0
	max	1.7 (±0.2)	5.0 (±0.5)	1.7 (±0.2)	17.0 (±0.4)	0.5	2.8 (±0.1)
YAPH	1/5	2.8	7.9 (±0.1)	1.9	14.8 (±0.2)	0.6	5.5 (±0.3)
	1/2	3.5	7.8 (±0.6)	2.5 (±0.2)	23.9 (±0.9)	1.0 (±0.3)	5.5
	max	3.1 (±0.1)	4.1 (±0.2)	1.7 (±0.2)	31.1 (±0.4)	1.5	4.5
YAHO	1/5	3.3 (±0.1)	3.6	1.9 (±0.2)	33.0	1.9 (±0.1)	3.4
	1/2	2.0	2.5 (±0.2)	2.3 (±0.1)	23.1 (±1.2)	1.2 (±0.1)	2.8 (±0.2)
	max	1.8 (±0.1)	2.0 (±0.4)	2.3 (±0.5)	23.3 (±3.8)	1.0 (±0.3)	2.4 (±0.2)
YAAL	1/5	2.2 (±0.2)	3.3 (±0.5)	1.9	7.9 (±2.2)	0.0	2.2
	1/2	3.1 (±0.1)	7.6 (±0.2)	4.5	19.4 (±0.2)	0.9	5.0
	max	1.3 (±0.2)	4.5 (±0.4)	4.5 (±0.1)	17.3 (±1.4)	0.5	3.6 (±0.1)
YAHI	1/5	5.5 (±0.3)	11.8 (±0.4)	1.5 (±0.1)	25.1 (±0.6)	0.8	6.7 (±0.3)
	1/2	6.6 (±0.1)	9.5 (±0.4)	1.7	35.8 (±0.1)	1.5	6.2 (±0.2)
	max	5.8 (±0.1)	7.3	1.3	38.3 (±2.8)	1.7 (±0.1)	6.1 (±0.1)
Glucose							
YALI	1/5	1.1 (±0.1)	0.7	0.1	1.6	0.0	2.3
	1/2	0.8	0.6	0.1	2.6 (±0.1)	0.0	1.4 (±0.1)
	max	0.7	0.7 (±0.1)	0.0	3.2 (±0.1)	0.0	1.3 (±0.1)
YADE	1/5	0.7 (±0.1)	0.4	0.0	1.6 (±0.3)	0.1	1.6 (±0.1)
	1/2	0.7 (±0.1)	0.4	0.0	2.1 (±0.4)	0.1	1.3 (±0.1)
	max	0.6 (±0.1)	1.1 (±0.2)	0.0	2.4 (±0.1)	0.0	1.3
YAGA	1/5	0.8	0.6	0.0	0.7	0.0	1.8 (±0.1)
	1/2	0.6 (±0.1)	0.5 (±0.1)	0.0	1.3 (±0.1)	0.0	1.4
	max	0.7	1.0	0.0	2.8 (±0.1)	0.0	1.2
YAYA	1/5	1.3 (±0.1)	1.4	1.7	0.7	0.0	2.3 (±0.1)
	1/2	1.1	1.4 (±0.1)	0.0	1.0 (±0.2)	0.0	1.7 (±0.1)
	max	0.6	1.2	0.1	1.3 (±0.1)	0.0	1.5
YAOS	1/5	0.5 (±0.1)	0.4 (±0.1)	0.0	1.3 (±0.4)	0.0	1.5 (±0.1)
	1/2	0.5	0.4	0.0	1.8 (±0.2)	0.0	1.3 (±0.2)
	max	0.9	0.9 (±0.1)	0.0	2.9 (±0.1)	0.0	0.8 (±0.1)
YAPH	1/5	1.1	0.4	0.0	1.5	0.2 (±0.1)	1.9
	1/2	0.9	0.4	0.0	2.4 (±0.2)	0.2	1.5 (±0.1)
	max	1.1	1.1 (±0.1)	0.0	3.3 (±0.3)	0.0	1.5 (±0.1)
YAHO	1/5	0.7	0.4 (±0.2)	0.0	1.2 (±0.4)	0.2	1.2 (±0.4)
	1/2	0.6	0.6	0.0	2.0 (±0.1)	0.0	1.3 (±0.1)
	max	0.6	1.0	0.0	3.2 (±0.1)	0.0	1.1 (±0.2)
YAAL	1/5	0.4	0.4 (±0.1)	0.1	0.7	0.0	2.3 (±0.1)
	1/2	0.4 (±0.1)	0.3 (±0.1)	0.1	1.0	0.0	2.4 (±0.1)
	max	0.4 (±0.1)	0.8	0.3	1.6 (±0.1)	0.0	2.0 (±0.2)
YAHI	1/5	1.4	0.7	0.0	0.9 (±0.2)	0.0	2.0 (±0.1)
	1/2	2.0	0.7 (±0.1)	0.0	1.9 (±0.1)	0.0	1.8 (±0.1)
	max	2.1 (±0.1)	0.0	1.1 (±0.1)	2.1 (±0.1)	0.1	1.8

CHAPITRE 2 :

LE MÉTABOLISME LIPIDIQUE DU CLADE *YARROWIA*

I. LES OBJECTIFS

L'objectif premier de cette partie de ma thèse était de retracer, d'un point de vue génomique, le métabolisme lipidique complet de YALI. En effet, lors de ma première année de thèse j'ai observé un manque de publications alliant à la fois des données protéiques et des listes de gènes utilisables par un génomiste. J'ai alors eu comme but de fournir à la fois un schéma des différentes voies de ce métabolisme et un tableau complet des gènes avec leur nomenclature, leur annotation et la référence bibliographique pour les enzymes déjà décrites. Sachant que certaines étapes de ce métabolisme n'avaient jamais été caractérisées chez YALI il a fallu proposer plusieurs gènes candidats qui coderaient des enzymes assurant les réactions manquantes. Contrairement à SACE, peu d'enzymes ont été caractérisées de manière expérimentale chez YALI. Dans ce contexte, la génomique comparative nous offre un outil puissant pour caractériser fonctionnellement des gènes pour lesquels aucune expérimentation n'a été réalisée. Une fois ce métabolisme reconstitué, le second objectif était de retrouver tous les gènes orthologues dans les génomes disponibles du clade *Yarrowia* et de OLHI. Pour cette partie je me suis alors investie dans l'annotation des 5 nouveaux génomes séquencés dans l'équipe, principalement pour les gènes impliqués dans le métabolisme des lipides.
La reconstitution du métabolisme lipidique chez YALI, dans le clade et chez OLHI permet alors de répondre à deux questions majeures :

- Les espèces du clade *Yarrowia* et OLHI possèdent-elles les mêmes voies constituant le métabolisme lipidique que l'espèce de référence YALI ?
- Toutes les fonctions de ce métabolisme sont elles assurées par des enzymes homologues dans chacune de ces espèces ?

II. INTRODUCTION

Ce chapitre présente les principales voies métaboliques et les compartiments cellulaires impliqués dans la dégradation, la synthèse et le stockage des substrats hydrophobes chez les levures. Il prend pour exemple YALI, quand des données sont publiées, ou à défaut SACE qui a une bibliographie très renseignée mais qui malheureusement n'est pas oléagineuse. De plus, les voies du métabolisme lipidique spécifiques aux levures oléagineuses sont présentées.
La reconstruction du métabolisme lipidique présentée est une combinaison des données issues de la bibliographie, de KEGG, d'UniProt et des annotations fonctionnelles des gènes de YALI et de SACE.

Les substrats hydrophobes (SH) tels que les alcanes, les acides gras et les triacyglycérides ont besoin d'entrer dans les cellules par l'intermédiaire de transporteurs qui ne sont pas encore connus chez YALI (Flèches noires sur la Figure). Pour l'import dans la cellule les triacyglycérides sont préférentiellement hydrolysés au préalable par des lipases. Lorsque le substrat est un alcane il est d'abord oxydé par les systèmes d'alcane monooxygénase P450 dans le réticulum endoplasmique et converti en alcool gras en deux étapes dans les peroxysomes ou dans le réticulum endoplasmique (flèches rouges sur la Figure 2.1, pour plus de détails voir paragraphe 1.2). Si le substrat est un acide gras il est activé, par l'acétyl-CoA synthétase ou l'acide gras activase, soit dans le réticulum endoplasmique soit dans les peroxysomes (flèches bordeaux sur la Figure 2 .1, pour plus de détails voir paragraphe 2.1) avant d'entrer dans la voie de ß-oxydation (flèches violette sur la Figure 2.1, pour plus de détails voir paragraphe 2.2). L'acétyl-CoA formé par la ß-oxydation peut être transporté dans les mitochondries sous forme d'acétyl-carnitine à l'aide de l'acétyl-carnitine acyl-transférase. L'acétyl-CoA peut également entrer dans le cycle du glyoxylate, situé dans les peroxysomes et interagir avec le cycle des acides tricarboxyliques et le cycle du méthylcitrate dans les mitochondries (flèches orange et jaunes sur la Figure 2.1, pour plus de détails voir paragraphe 3.3.2). Les acides gras peuvent aussi être stockés dans des corps lipidiques sous forme de triacyglycérides ou d'esters de stérol (compartiment orange sur la Figure 2.1, pour plus de détails voir paragraphe 6).

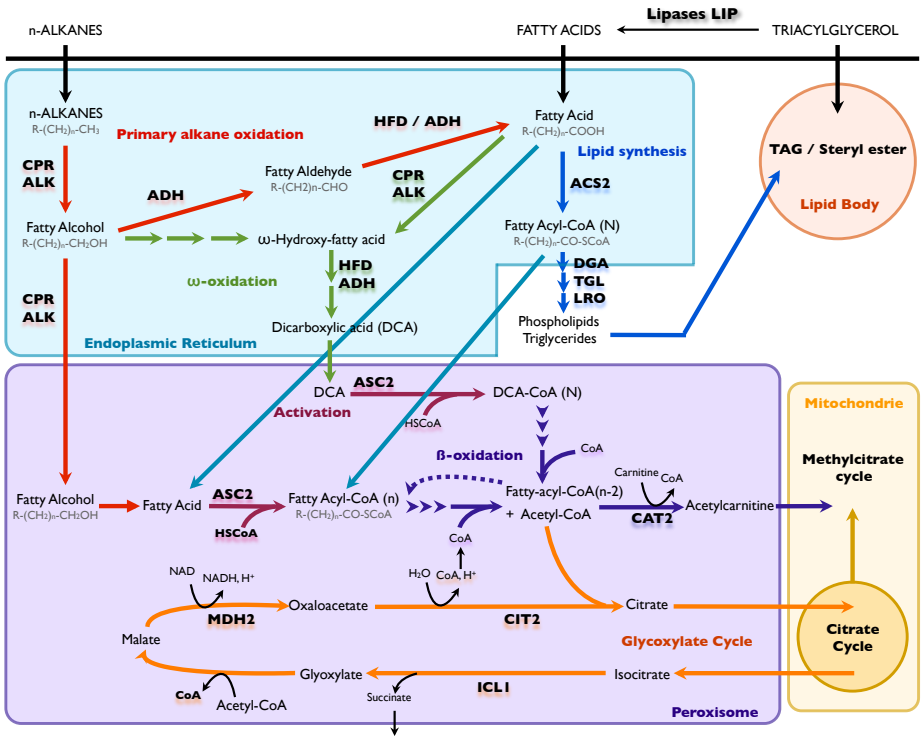


Figure 2.1 : Schéma des principales voies métaboliques impliquées dans l'assimilation et la dégradation des substrats hydrophobes. Les principaux flux métaboliques pour la synthèse des lipides sont représentés en bleu, pour l'oxydation primaire des alcanes en rouge, pour l'ω-oxydation en vert, pour l'activation des acides gras en bordeaux, pour la ß-oxydation en violet et pour le cycle du glyoxylate en orange. L'import des substrats hydrophobes est symbolisé par les flèches noires (modifié pour YALI à partir de Fickers et al., 2005a).

Quelques définitions biochimiques ...

Aldéhydes gras : acides gras dont le groupement carboxyle -COOH a été substitué par un groupement aldéhyde -CH=O.

Alcanes : hydrocarbures saturés.

Acides gras : acides carboxyliques caractérisés par une répétition de groupements méthyles -CH2- formant une longue chaîne carbonée généralement constituée d'un nombre pair d'atomes de carbone.

Alcools gras : acides gras dont le groupement carboxyle -COOH a été substitué par un groupement hydroxyle –OH.

Sucres : molécules contenant un groupement carbonyle (aldéhyde ou cétone) et plusieurs groupements hydroxyle –OH.

1 IMPORT DES LIPIDES

1.1 Hydrolyse des triglycérides : les lipases

Les lipases, sous-classe des estérases, sont des enzymes capables de catalyser les réactions de synthèse et d'hydrolyse des esters de stérols et des triglycérides. Toutes les lipases possèdent une triade catalytique avec la sérine, l'acide aspartique et l'histidine conservés caractéristiques de la triade hydrolase α / β et le nucléophile catalytique situé dans le pentapeptide hautement conservé GHSXG (Schrag et al., 1997). En raison de leur polyvalence les lipases sont largement utilisées dans diverses applications industrielles. La levure modèle SACE ne possède pas de lipase et est incapable d'utiliser les TAG à partir du milieu externe. Toutefois, dans le génome de SACE on trouve par blastp une « putative lipase », YJR107W LIH1 qui possède un motif GHSLG et la triade catalytique (Figure 2.2). Les études sur les lipases s'intéressent donc à des levures non conventionnelles telles que *C. rugosa* qui possède une famille de 5 lipases (Benjamin and Pandey, 1998), *C. deformans* possédant au moins trois lipases dont une extracellulaire (Bigey et al., 2003), *C. albicans* avec une grande famille comprenant au minimum 10 lipases pouvant contribuer à la virulence de cette levure dans les cellules humaines (Hube et al., 2000) et plus récemment *C. viswanathii* possédant une lipase efficace pour l'hydrolyse d'une large gamme de triglycérides naturels (de Almeida et al., 2013).

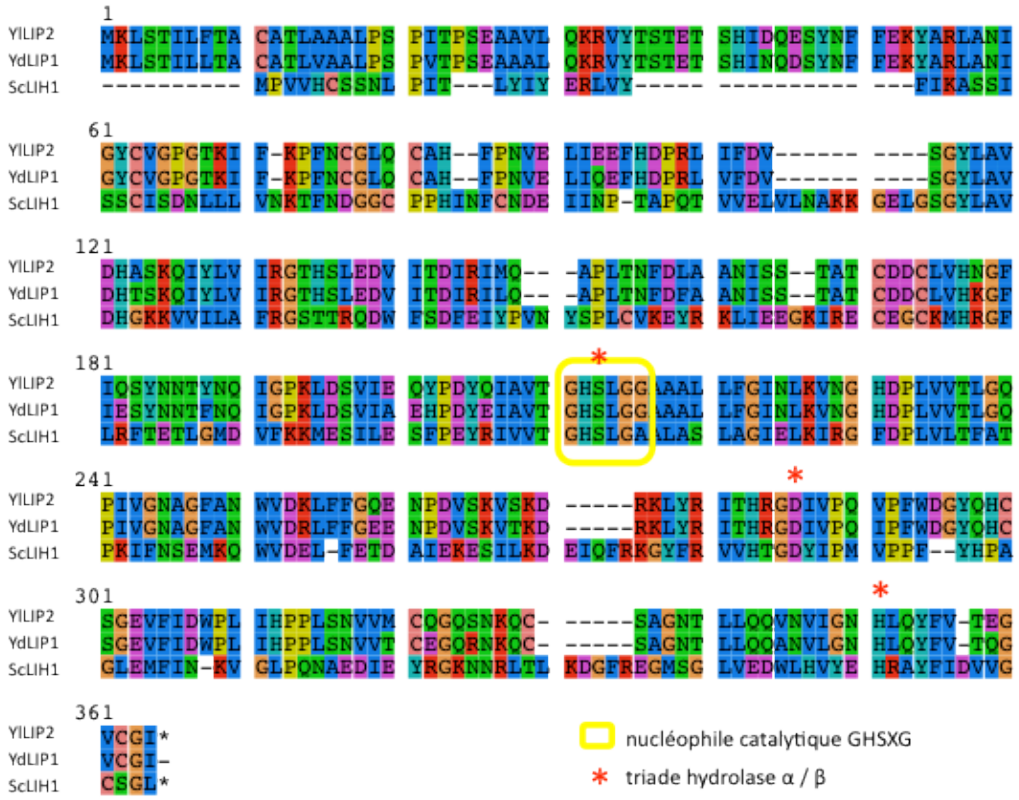


Figure 2.2 : Alignement des protéines YlLIP2, YdLIP1 et ScLIH1. Le motif GHSXG est entouré en jaune et la triade catalytique S-D-H est marquée par des étoiles rouges.

Mais la plupart des études sur les lipases de levures sont menées chez YALI. La particularité de YALI est de posséder 16 gènes paralogues codant des lipases incluant des enzymes intracellulaires, liées à la membrane et extra-cellulaires. Elles possèdent au minimum 31,4% de similarité avec ScLIH1 pour la YlLIP11 et au maximum 43,8% pour YlLIP2 (Figure 2.2). Avec ces 16 lipases elle est capable d'utiliser de nombreux substrats hydrophobes. Pignède *et al.* ont montré que l'activité lipasique extracellulaire résulte de l'expression de YlLIP2 (Pignède et al., 2000). YlLIP2 a donc été reconnue comme très prometteuse au niveau industriel et suscite de nombreuses études. En 2011, Bordes *et al.* ont décrit la thermosensibilité d'un variant de LIP2 (Bordes et al.), Darvishi *et al.* ont réussi à produire une grande quantité de LIP2 par mutation induite (Darvishi et al., 2011) et Shockey *et al.* ont cloné YlLIP2 dans SACE sous le promoteur PEX11 de SACE permettant ainsi à la souche de sécréter l'enzyme dans le milieu en réponse à la présence de lipides (Shockey et al., 2011). En 2012, le gène LIP2 natif et un mutant de LIP2 son été exprimés dans SACE (Darvishi, 2012), cette souche est alors capable d'utiliser des lipides et de l'huile d'olive pour produire du bioéthanol et d'autres produits d'intérêt biotechnologique. Actuellement cette lipase est encore très étudiée et déjà utilisée pour de nombreuses applications biotechnologiques telles que le traitement des eaux usées et plus particulièrement de celles provenant des moulins à huile ; dans le domaine pharmaceutique pour la synthèse de médicaments et pour le traitement de l'insuffisance pancréatique (pour revue voir (Fickers et al., 2005a; Fickers et al., 2011).

Les autres lipases de YALI ont été moins étudiées. Les gènes LIP7 et LIP8 ont été isolés par homologie avec les gènes LIP2 et LIP3 de *C.deformans*. Ils codent des lipases fonctionnelles et partiellement sécrétées (Fickers et al., 2005b). *In vitro*, LIP7 a une activité optimale pour les ester à chaine carbonée de taille moyenne (C6) alors que LIP8 montre une meilleure activité pour les ester en C8 et C10. Leur localisation à la surface des cellules a été confirmée par Liu *et al.* (Liu et al., 2010). Puis LIP16 a été isolée. Elle serait spécifique des substrats en C10 *in vitro* (Zhao et al., 2011b). En 2012, deux nouvelles lipases extracellulaires et fonctionnelles ont été étudiées : LIP14 et LIP18 montrant une forte préférence *in vitro* pour les TAG et huiles riches en acide oléique (Kumari and Gupta, 2012). La même année les enzymes LIP11 et LIP12 ont également été caractérisés comme des enzymes extracellulaires. LIP11 est proche de LIP2, avec une spécificité *in vitro* pour les substrats à moyenne et longue chaine alors que LIP12 a seulement une très faible spécificité pour les chaines courtes à moyennes et une faible activité catalytique (Kumari and Gupta, 2013; Kumari et al., 2012). Les lipases LIP4 et LIP5 semblent avoir été également étudiées mais les articles ne sont pas accessibles car ils sont publiés en chinois (Zhang et al., 2010; Zhao et al., 2011a). Ainsi, sur les 16 lipases de YALI seulement huit ont été caractérisées au niveau biochimique : LIP2, LIP7, LIP8, LIP9, LIP11, LIP12, LIP14 et LIP18 et de manière générale, la spécificité, la régulation et l'expression des LIP sont très peu connues.

1.2 Multiples voies métaboliques pour les substrats hydrophobes

1.2.1 Des alcanes aux acides gras

Contrairement à SACE, YALI est capable d'utiliser les alcanes comme seule source de carbone. Selon Thevenieau *et al.* chez YALI, les protéines codées par ABC1 YALI0E14729g permettent l'importation des alcanes en C14 à C16 dans le réticulum endoplasmique (RE) et celles codées par ABC2 YALI0C20265g, ABC3 YALI0B02544g et ABC4 YALI0B12980g celle des alcanes en C10 à C12 (Thevenieau et al., 2007). Selon les études transcriptomiques de Chen *et al.* le niveau de mRNA de ABC2 augmente significativement lorsque YALI est en présence d'octane (C8), nonane (C9), décane (C10) and undécane (C11) et le niveau de mRNA de ABC3 augmente significativement en présence de nonane et de décane (Chen et al., 2013). Ceci suggère que ABC2 aurait une spécificité pour les alcanes en C8 à C11 et ABC3 pour les C9 et C10. Par contre ABC1 et ABC4 ne montrent pas de changement dans leur expression en fonction des différents alcanes testés. Ces auteurs confirment également la localisation de ABC2 et ABC3 dans les membranes plasmiques (Chen et al., 2013). Chez SACE, bien qu'elle ne soit pas capable de pousser sur alcane, il existe également deux protéines SNQ2 et PDR5, codées respectivement par YDR011W et YOR153W et homologues à YlABC5 YALI0F17996g et YlABC3 ayant un potentiel pour l'efflux des alcanes ce qui contribuerait à la capacité à tolérer les alcanes dans la cellule (Ling et al., 2013).

Chez YALI, après leur import dans la cellule, si les alcanes ne sont pas en quantité toxique, ils vont entrer dans la voie de l'oxydation primaire des alcanes. Dans cette voie ces substrats sont oxydés en alcools gras de même longueur de chaine dans le réticulum endoplasmique (RE) par les cytochromes P450 codés par YlALK1 à YlALK12 associés à la NADPH-cytochrome P450 reductase YlCPR1 YALI0D04422g. Selon Iida *et al.* ALK1 catalyse l'oxydation pour une grande variété d'alcane et plus spécialement les n-decanes, ALK2 serait spécialisé pour les substrats plus longs tel que le n-hexadécane et les autres ALK agiraient sur les alcanes plus longs ou sur d'autres types de chaines hydrocarbonées (Iida et al., 1998; Iida et al., 2000)

Les alcools gras sont alors oxydés pour donner des aldéhydes gras de même longueur de chaine. D'après Fickers *et al.,* chez les levures, cette oxydation est réalisée soit directement dans le RE par des alcools déshydrogénases ADH, soit les alcools gras sont d'abord exportés dans le peroxysome pour y être oxydés par des alcools gras oxydases FAO (Fickers et al., 2005a). SACE ne possède probablement par de FAO mais elle a 12 gènes codant des ADH (SFA1, SOR1, SOR2, XYL2, BDH1, BDH2, ADH1, ADH2, ADH3, ADH5, ADH6 et ADH7). Parmi ces gènes, ADH6 et ADH7 sont semblables aux enzymes de plantes permettant le passage de l'aldéhyde à l'alcool dans la voie de la dégradation de la lignine (Larroy et al., 2003). SACE est en effet capable d'utiliser des fractions dérivées de lignine comme seule source de carbone, elle peut donc croître dans un environnement ligninolytique si d'autres microorganismes ont réalisé une première dégradation

de la lignine (Kosíková and Sláviková, 1996). Chez YALI deux ADH NAD+ dépendante et trois NADP+ dépendante ont été détectées par Barth *et al.* par électrophorèse sur gel de polyacrylamide (Barth and Gaillardin, 1996; Barth and Künkel, 1979) mais nous ne trouvons aucune donnée génétique sur les ADH ni sur une potentielle FAO chez YALI.

Les aldéhydes gras dans le peroxysome ou le RE donnent ensuite des acides gras (AG) de la même longueur de chaine grâce aux aldéhydes gras déshydrogénases HFD1 à HFD4 codées par les gènes YlHFD1 à YlHFD4 (YALI0B01298g, YALI0F23793g, YALI0A17875g, YALI0E15400g) (Matatiele, 2005). Ces composés peuvent sortir du RE, sous forme d'alcools ou d'AG pour aller dans le peroxysome. Les AG sont ensuite activés par thio-estérification au coenzyme A par des Acyl-CoA synthétases pour être utilisés dans les différentes voies métaboliques.

1.2.2 La création des acides dicarboxyliques (DCA)

La voie de l'ω-oxydation produit des acides dicarboxyliques (DCA) à partir des AG et ses alcools gras par les actions successives des cytochromes P450 (ALK1 à ALK12) combinés à la NADPH-cytochrome P450 reductase CPR puis des alcools déshydrogénases ADH et des aldéhydes gras déshydrogénases HFD1 à HFD4. Les DCA seront alors activés comme les AG pour rentrer dans la ß-oxydation peroxysomale.

1.3 Transport et activation des acides gras dans le cytosol

Les AG peuvent être directement importés du milieu extérieur. Même si YALI est capable de les synthétiser *de novo* cela permet d'économiser l'énergie cellulaire dépensée lors des réactions de synthèse. YALI peut incorporer les AG sous leur forme libre mais pour être utilisés dans les différentes voies métaboliques ils sont le plus souvent activés par thio-estérification au coenzyme-A (CoA) par des Acétyl-CoA synthétases. Cette réaction correspond à la formation d'une liaison thioester entre le carboxyle de l'acide gras et le groupe thiol du coenzyme A. Ces étapes peuvent être réalisées par un processus d'entrée-activation qui combine l'entrée des AG dans la cellule et son activation. Ce mécanisme est connu chez SACE (Choi and Martin, 1999; Obermeyer et al., 2007), il est réalisé par les protéines Fat1p et Faa1p ou Faa4p. Fat1p, codées par ScFAT1 YBR041W, permet l'importation des AG. Sa délétion implique une diminution de 40 à 70% de l'activation des AG selon leur longueur de chaine et une accumulation des acides gras à très longue chaine carbonée (VLCFA) (>20C) car ils ne peuvent plus être dégradés. En effet, les Faa1p et Faa4p, codées respectivement par ScFAA1 et ScFAA4, sont des acyl-CoA synthétases qui interagissent directement avec Fat1p. Ils permettent l'activation des AG importés en utilisant de l'ATP + CoA, avec une spécificité pour les substrats à très longue chaine carbonée VLCFA (Zou et al., 2003). Ce mécanisme est appelé l'acylation vectorielle membranaire (Figure 2.3).

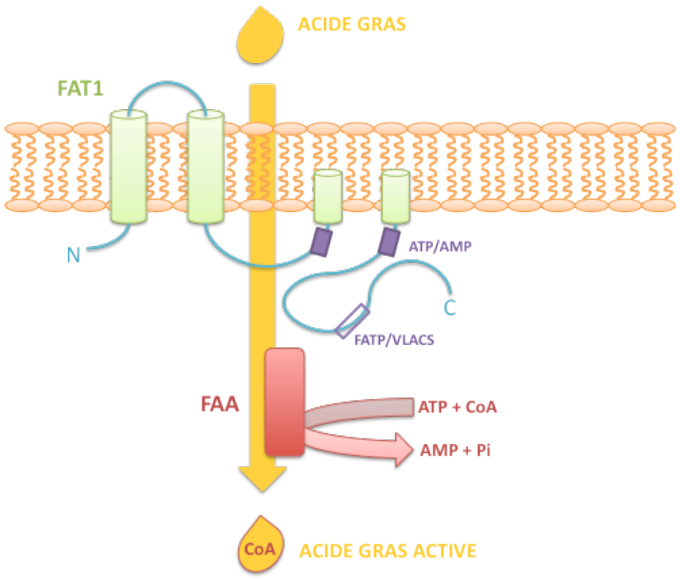


Figure 2.3 : Modèle topologique de FAT1 et des FAA chez SACE. Pour la FAT1 les segments transmembranaires sont notés comme des cylindres. Les positions approximatives des motifs de l'ATP/AMP et FATP/VLACS (très longue chaîne acyl-CoA synthétase) sont notées respectivement par les rectangles fermés et ouverts (adapté de (Obermeyer et al., 2007).

2 Dégradation des acides gras activés dans les peroxysomes

2.1 Activation et transport des acides gras dans les peroxysomes

Les peroxysomes sont des vésicules entourées d'une membrane unique composée d'une bicouche lipidique. Ils ne possèdent ni matériel génétique ni machinerie traductionnelle et sont donc obligés d'importer l'ensemble de leurs éléments à partir du cytoplasme. Pour pouvoir entrer dans les peroxysomes, une protéine doit avoir un signal d'adressage spécifique. Cet adressage se fait par des séquences signal appelées « peroxisomal targeting signals » (PTS). Il existe un PTS1 en C-terminal composé de trois peptides consensus : (S/A/C)-(K/R/H)-(L/M) (Rachubinski and Subramani, 1995). Il existe également un PTS2 en N-terminal. La séquence consensus du PTS2 généralement acceptée est (R/K)-(L/V/I)-xxxxx-(H/Q)-(L/A) (Lazarow, 2006). Toutes les protéines ayant un rôle dans les peroxysomes doivent donc avoir un PTS excepté quelques polypeptides qui peuvent entrer dans les peroxysomes en s'associant à une sous unité possédant un PTS (Glover et al., 1994). Ces deux séquences sont mutuellement exclusives, et PTS1 est la séquence d'adressage la plus souvent retrouvée.

Chez SACE les AG peuvent entrer dans les peroxysomes de deux façons (Figure 2.4). Les acides gras de longue à très longue chaine carbonée (VLCFA et LCFA) sont activés dans le cytoplasme et rentrent dans les peroxysomes grâce à l'hétérodimère composé de Pxa1p et Pxa2p, appartenant à la famille des ABC transporteurs, codés par YKL188C et YPL147W (Hettema and Tabak, 2000;

Verleur et al., 1997). Selon van Roermund *et al.* en 2000, les acides gras à chaine carbonée de taille moyenne (MCFA) entrent comme des AG libres en utilisant Pex11p et l'acyl-CoA synthétase Faa2p YOR317W pour leur activation dans les peroxysomes (van Roermund et al., 2000). Le même auteur en 2012 ne parle plus de Pex11 mais rajoute Fat1p YBR041W sur la membrane externe du peroxysome pour le transport des AG (van Roermund et al.). Fat1p et les Faa2p agissent alors pour l'entrée et l'activation dans les peroxysomes de la même façon que Fat1p et Faa1p ou Faa4p dans le cytosol. L'activation par Faa2p nécessite un apport d'ATP permis par Ant1p, codé par ScANT1 YPR128C, qui catalyse les échanges ATP-AMP à travers la membrane peroxysomale (Palmieri et al., 2001).

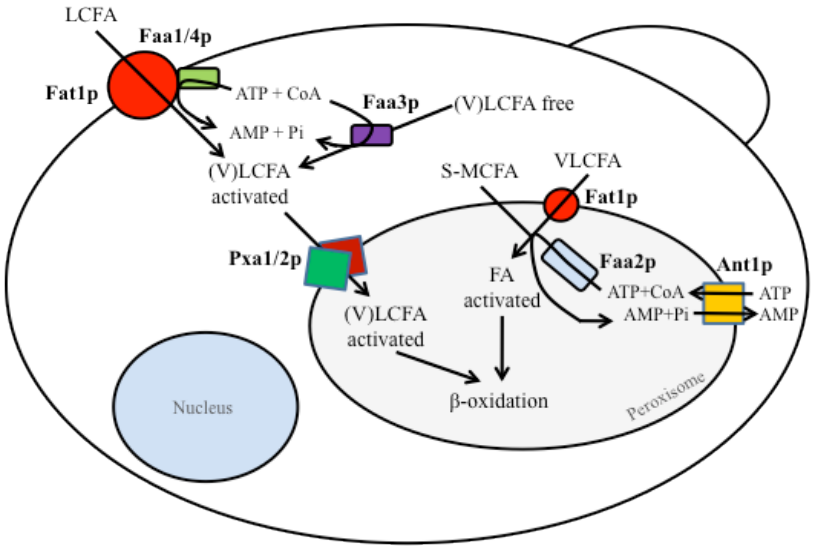


Figure 2.4 : Schéma des principaux mécanismes de transport des AG à travers les membranes chez SACE. VLCFA : AG à très longue chaine carbonée > 22C ; LCFA : AG de 13C à 21C ; MCFA : AG de 8C à 12C; SCFA: AG < à 8C ; Pxa1p et Pxa2p : ABC transporteurs peroxysomaux; Fat1p : fatty acid transporteur ; Faa1p à Faa4p : l'acyl-CoA synthétase ; Ant1p : Peroxisomal adenine nucleotide transporter (schéma d'après R. Dulermo, communication personnelle)

Thevenieau *et al.* suggèrent que chez YALI les MCFA entrent dans les peroxysomes sous forme non activée grâce à un transporteur non identifié et que les LCFA entreraient aussi sous forme non activée mais avec le transporteur PXA1/PXA2, codé par les gènes YlPXA1 YALI0A06655g et YlPXA2 YALI0D04246g homologues à ScPXA1 YPL147W et ScPXA2 YKL188C. Les FA importés seraient ensuite activés dans les peroxysomes par l'acétyl-CoA synthétase ACS2 (Thevenieau et al., 2007). Les homologues des gènes codant Pex11p et Ant1p existent chez YALI suggérant que l'activation simultanée au transport des AG est aussi possible.

2.2 Bêta oxydation : un cycle de réactions et plusieurs compartiments

2.2.1 Plusieurs compartiments

La voie majeure de dégradation des acides gras saturés ou insaturés est la ß-oxydation. Dans la majorité des cellules eucaryotes elle se déroule à la fois dans les peroxysomes et dans les mitochondries, avec une spécialisation de la ß-oxydation mitochondriale pour les AG à chaine courte. Mais chez les levures et les plantes, la ß-oxydation peroxysomale assure aussi la dégradation des AG à chaîne courte permettant ainsi de dégrader entièrement les acides gras dans les peroxysomes.

2.2.2 Le principe

Dans les deux cas, la ß-oxydation catalyse le raccourcissement des acyl-CoA entre les carbones 2 et 3 ce qui donne un acyl-CoA à n-2 carbones avec un acétyl-CoA, ou un propionyl-CoA dans le cas des acyl-CoA à chaine impaire. La première étape de la ß-oxydation est la déshydrogénation des acyl-CoA entre les carbones 2 et 3 par une oxydo-réductase. L'oxydo-réductase est une acyl-CoA déshydrogénase dans les mitochondries et une acyl-CoA oxydase dans les peroxysomes. La seconde étape est l'hydratation de la double liaison, catalysée par une crotonase appelée enoyl-CoA hydratase, qui aboutit à un ß-hydroxyacyl-CoA. Ensuite il y a coupure de la chaine carbonée par une transférase formant un acétyl-CoA ou un propionyl-CoA et un acyl-CoA avec 2 carbones en moins qui repart dans le cycle de la ß-oxydation. L'oxydation des AG saturés et insaturés dans les peroxysomes ne requière pas seulement la participation d'une série d'enzymes. En effet les peroxysomes ont besoin de mécanismes pour importer les AG à partir du cytosol, pour réduire le $NADP^+$ produit dans la voie des AG insaturés, pour exporter l'acétyl-CoA hors du peroxysome et pour réoxyder le NADH produit durant la ß-oxydation.

2.2.3 Première étape de la ß-oxydation peroxysomale

La première étape de la ß-oxydation des AG activés est la création d'une double liaison entre les carbones 2 et 3. Chez SACE il y a une seule acyl-CoA oxydase localisée dans la matrice peroxysomale, ScPox1p, codée par YGL205w ScPOX1. Cette enzyme a une double régulation post traductionnelle avec la répression par le glucose et l'induction par l'acide oléique (Dmochowska et al., 1990). ScPox1p accepte toutes les longueurs de chaine et permet donc la dégradation des AG jusqu'à la simple unité d'acétyl-CoA à 2 carbones.

Le génome de YALI contient une famille de 6 gènes POX. Cinq avaient été identifiés par Wang *et al.* et le sixième a été trouvé par l'analyse du génome (Wang et al., 1999). Ce nombre est une

originalité chez les levures. Chez *C.maltosa* deux gènes POX ont été identifiés et trois chez *C.tropicalis* (Koei et al., 1987; Masuda et al., 1995). Chez YALI il y a une spécialisation de POX2 et POX3 en fonction de la longueur de chaine carbonée des substrats et les autres sont différenciées par leur niveau d'activité (Mlíčková et al., 2004) (Figure 2.5). En effet POX2 a une forte activité mais seulement pour les acyl-CoA à chaine longue (Luo et al., 2002). POX3 a une faible activité pour les acyl-CoA à chaine longue et une forte activité pour les acyl-CoA à chaine courte (Luo et al., 2000). POX4 et POX5 ont une activité moyenne et non spécifique. Enfin, POX1 et POX6 ont une très faible activité sans spécificité.

POX4 et POX5 semblent aussi intervenir dans la régulation de la division des peroxysomes en interagissant avec une protéine peroxysomale membranaire, PEX16 qui empêche la fission des membranes des peroxysomes immatures. Lorsque les peroxysomes sont matures, cette interaction entraine alors leur division. La présence de grands peroxysomes chez les mutants ΔPOX4 et ΔPOX5 de YALI confirme cette hypothèse (Guo et al., 2003).

Une souche mutante de YALI exprimant une polyhydroxyalkanoate (PHA) synthétase de *Pseudomonas aeruginosa* et chez qui la seconde étape de la ß-oxydation peroxysomale est bloquée peut synthétiser des PHA. Celle ci libère des acides gras hydroxylés en position 3 (3-hydroxyalcanoates) par l'action des POX durant la première étape de la ß-oxydation. Ces AG peuvent alors servir de substrat à une PHA synthétase. Dans cette souche productrice de PHA, POX3 seule serait responsable de 90% et 75% de la production totale de PHA à partir d'AG en C9 et C13, respectivement, et POX5 seule est responsable de 70% de la production de PHA à partir d'AG en C9 (Haddouche et al.).

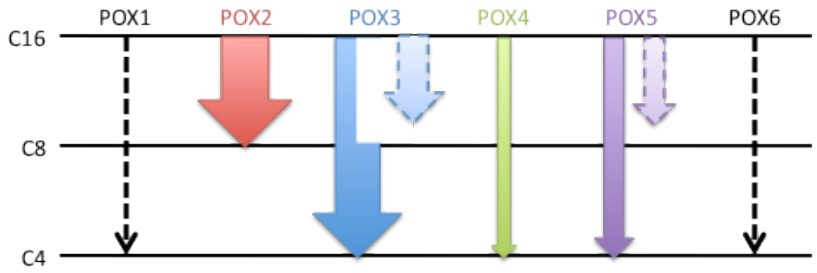


Figure 2.5 : Représentation schématique de l'activité relative et des spécificités des enzymes POX de YALI. Le niveau d'activité est indiqué par la largeur des flèches et la longueur indique la spécificité de longueur de chaîne du substrat (C16, C8, ou C4). Les flèches noires en pointillés représentent une très faible activité sans spécificité. Les flèches en pointillés colorées représentent l'activité et la spécificité pour la production des PHA chez une souche productrice de PHA (Haddouche et al., 2010). Figure modifiée de Mlíčková *et al.* (Mlíčková et al., 2004).

2.2.4 Deuxième et troisième étapes de la ß-oxydation peroxysomale

Les deuxième et troisième étapes sont catalysées par une enzyme multifonctionnelle MFE. Ses activités 2-enoyl-CoA hydratase et 3-hydroxyacyl-CoA déshydrogénase aboutissent à la formation du 3-cétoacyl-CoA à partir du trans-2-enoyl-CoA. Chez SACE elle est codée par MFE2 YKR009C qui contient un domaine PTS1 (Hiltunen et al., 1992).
Chez YALI la MFE est également une enzyme bifonctionnelle, elle est codée par YlMFE ayant un PTS1. (Haddouche et al., 2011).

2.2.5 Dernière étape de la ß-oxydation peroxysomale

La dernière étape est catalysée chez SACE par une seule 3-cetoacyl-CoA thiolase, ScPot1p, qui clive le 3-hydroxyacyl-CoA en l'acyl-CoA correspondant à 2 carbones en moins et un acétyl-CoA. Chez SACE l'expression du gène codant la thiolase, POT1 YIL160C, est induite par les AG (Einerhand et al., 1991). Sa séquence contient un motif PTS2 (MSQ**RLQSIKDHL**) lui permettant d'être adressé au peroxysome.
YALI possède deux thiolases : YlPOT1 (YALI0E18568g) et YlPAT1 (YALI0E11099g). En 1993, Berninger *et al.* observent que la disruption de POT1 abolit la capacité de YALI à croitre sur acide oléique (AO). Ils suggèrent alors que POT1 est une 3-cétoacyl-CoA thiolase peroxysomale et qu'il n'y en a pas d'autre (Berninger et al., 1993). Mais en 2001, il a été montré que YlPAT1 code une acétoacétyl-CoA thiolase peroxysomale avec un domaine proche du domaine consensus PTS2 (M**RL**TLPRL**NA**) et qu'elle est essentielle pour la dégradation des *n*-décanes (C10) (Yamagami et al., 2001).

2.2.6 La ß-oxydation mitochondriale

Nous avons vu précédemment que POX3 est spécifique des AG à chaines carbonées courtes et que POX4 et POX5 sont aussi capables de les accepter. C'est donc grâce à ces POX que la ß-oxydation mitochondriale, spécifique des chaînes courtes, n'est pas nécessaire chez YALI. Cependant après la délétion des gènes essentiels de la ß-oxydation peroxysomale, Haddouche *et al.* observent une ß-oxydation résiduelle. Ils suggèrent alors l'existence d'une ß-oxydation mitochondriale (Haddouche et al., 2011). Vorapreeda *et al.* proposent que la première étape soit réalisée par une acyl-CoA déshydrogénase codée soit par YlACD2 ou YlACD3 ou YlACD4. Les trois autres étapes mettraient en jeu la 3-hydroxyacyl-CoA déshydrogénase puis l'enoyl-CoA déshydratase et l'acyl-CoA acetyltransferase codées respectivement par YALI0C08811g, YlECH2 et YlPAT1 (Vorapreeda et al., 2012). Cette voie, considérée classiquement comme absente chez les levures, rapprocherait YALI des champignons filamenteux.

2.2.7 La ß-oxydation des acides gras insaturés : les enzymes auxiliaires

Les étapes classiques de la ß-oxydation sont suffisantes pour les AG saturés et ceux ayant des doubles liaisons trans en position paire ainsi que les acides gras 2S-méthyl ramifié si une acyl-CoA oxydase méthyl-ramifié est présente. Les AG cis- et trans- insaturés ayant des doubles liaisons en position impaire et les AG cis- insaturés en position paire peuvent également être utilisés comme source de carbone mais doivent subir des étapes supplémentaires.

Chez SACE, les AG *cis*- avec une double liaison en position paire de type *cis*-4-enoyl-CoA ne peuvent rentrer dans la ß-oxydation qu'après leur conversion en 2,4-dienoyl-CoA par l'acyl-CoA oxydase ScPOX1 (première étape de la ß-oxydation), puis en trans-2-enoyl-CoA par l'action successive des enzymes 2,4-dienoyl-CoA réductase SPS19 (YNL202W) et Δ^3,Δ^2-enoyl-CoA isomérase ECI1 (YLR284C) (Figure 2.6). Les AG avec une double liaison en position impaire de type 2,5-dienoyl-CoA peuvent suivre deux routes. Soit ils sont convertis en trans-2-enoyl-CoA par le cycle classique de la ß-oxydation puis par l'action de l'ECI1 ; soit ils sont convertis en trans-2-enoyl-CoA grâce aux actions successives des enzyme : ECI1, une $\Delta^{3,5}$,$\Delta^{2,4}$-dienoyl-CoA isomérase DCI1 (YOR180C), SPS19 et de nouveau ECI1. Dans les trois cas, les trans-2-enoyl-CoA générés par l'ECI1 partent ensuite dans la ß-oxydation classique (Hiltunen et al., 2003). Chez YALI la ß-oxydation des acides insaturés est peu connue. L'enzyme 2,4-dienoyl-CoA réductase SPS19 a été purifiée par Mizugaki *et al.* en 1985 (Mizugaki et al., 1985). Ces auteurs suggèrent que cette enzyme est un polymère composé de dix à douze sous unités identiques et qu'elle n'est présente qu'en une seule copie chez YALI. Les autres enzymes, ECI1 et DCI1, n'ont pas été identifiées jusqu'ici chez YALI.

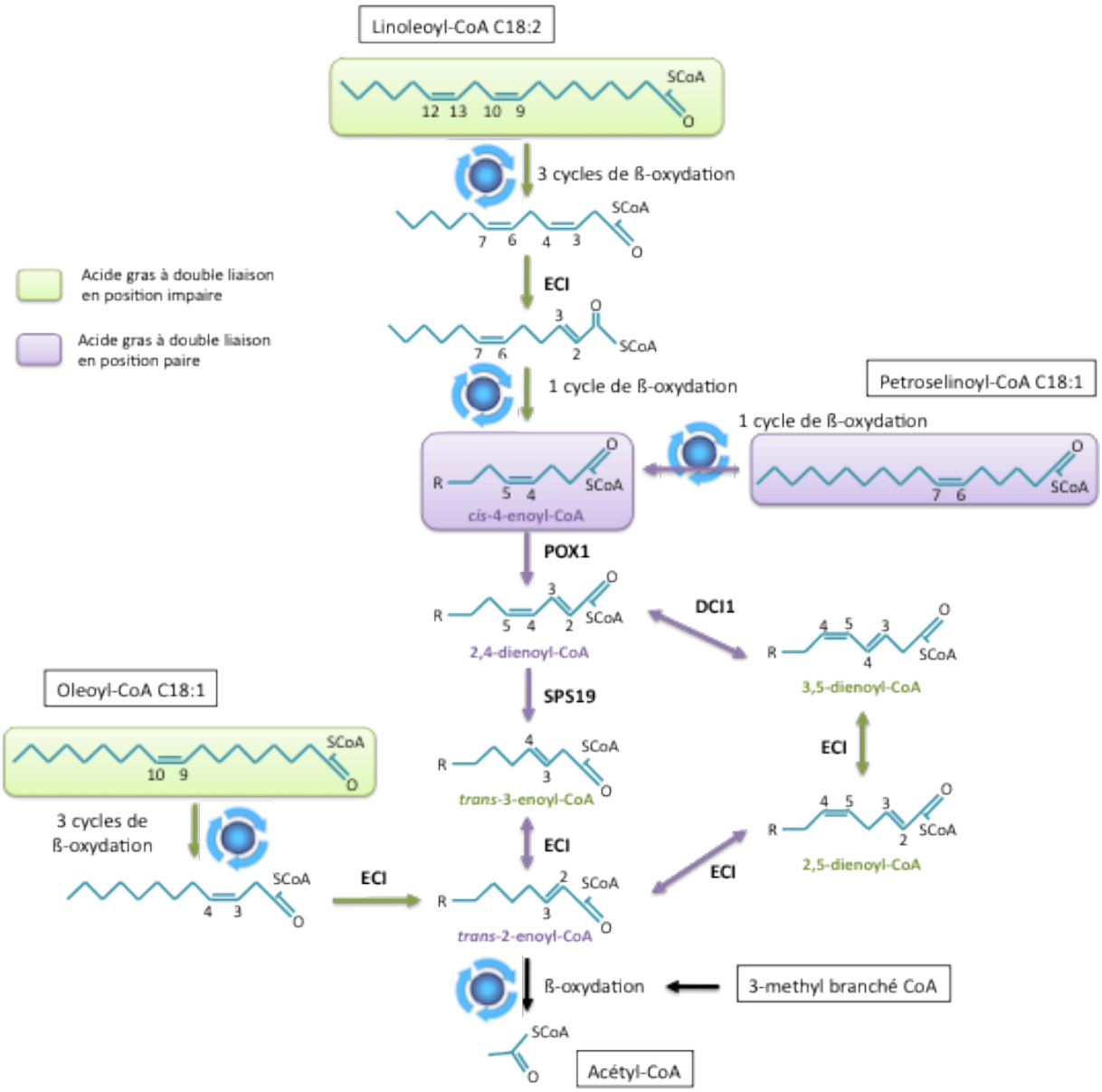


Figure 2.6 : Les étapes requises pour la ß-oxydation des acides gras insaturés**.** La dégradation des acides gras *cis*- avec une double liaison en position paire (cadres violets) est représentée avec des flèches violettes. Un exemple est donné avec le petroselinoyl-CoA, C18:1 de type *cis*- (cadre violet). Les deux voies de dégradation des acides gras ayant une double liaison en position impaire (cadre vert) sont représentées avec des flèches vertes. Un exemple est donné avec l'oleoyl-CoA, C18:1 de type *trans*- , et avec le linoleoyl-CoA, C18:2 de type *trans*- (cadres verts) . Les noms des AG avec une double liaison en position paire sont notés en violet et ceux des AG avec une double liaison en position impaire en vert. SPS19 : 2,4-dienoyl-CoA reductase ; ECI1 : Δ^3,Δ^2-enoyl-CoA isomérase ; DCI1 : $\Delta^{3,5}$,$\Delta^{2,4}$-dienoyl-CoA isomérase ; POX1 : acyl-CoA oxydase. Adaptée de van Roermund *et al.* et Poirier *et al.* (van Roermund et al., 1998).

2.2.8 Cas des acides gras à nombre impair d'atomes de carbone

La ß-oxydation peut également dégrader des AG avec un nombre impair de carbone. Dans ce cas, le produit de chaque cycle n'est pas un acétyl-CoA mais un acétyl-CoA plus un propionyl-CoA. Le propionyl-CoA ne peut pas entrer dans la voie du methylmalonyl-CoA, traditionnellement connue chez les procaryotes, car l'enzyme clé de cette voie, la methylmalonyl-CoA mutase, nécessite un apport de vitamine B12 que les levures ne synthétisent pas (Pronk et al., 1994). Une autre voie a donc été mise en évidence chez YALI puis SACE. Dans cette seconde voie, le propionyl-CoA entre dans le cycle du 2-methylcitrate où il est oxydé par la methylcitrate synthase, codée par YlCIT2, pour fournir du pyruvate (Tabuchi and Uchiyama, 1975; Uchiyama and Tabuchi, 1976). La methylcitrate synthase jouerait également le rôle de régulatrice pour ce cycle (Uchiyama et al., 1982; Uchiyama and Tabuchi, 1976).

2.2.9 L'acyl-CoA thioesterase

Chez SACE l'acyl-CoA thioesterase ScTES1p (YJR019C) permet de réguler les quantités d'acyl-CoA, d'AG libre et de CoASH en hydrolysant l'acyl-CoA en AG libres et CoA (Maeda et al., 2006). Cette réaction n'est pas caractérisée jusqu'ici chez YALI.

3 L'ACÉTYL-COA

3.1 La synthèse de l'acétyl-CoA

Nous avons vu que l'acétyl-CoA était synthétisé par la ß-oxydation peroxysomale ou mitochondriale lorsqu'elle existe. Mais il est également synthétisé par d'autres voies qui sont la dégradation de la leucine, la synthèse à partir du citrate ou de l'acétate par les acétyl-CoA synthétases et par la dégradation du pyruvate.

3.1.1 Du pyruvate à l'acétyl-CoA

Quand la source de carbone disponible pour la cellule est glycolytique le pyruvate est la principale source d'acétyl-CoA. Le pyruvate subit alors une décarboxylation dans la mitochondrie pour donner de l'acétyl-CoA qui peut être utilisé par le cycle du TCA par exemple. Pour entrer dans la mitochondrie le pyruvate utilise un transporteur récemment identifié chez la levure, composé de deux protéines chez SACE codées par les gènes respectifs ScMPC1 et ScMPC2 (Bricker et al., 2012). Puis la décarboxylation est catalysée par le complexe pyruvate déshydrogénase. Le complexe pyruvate déshydrogénase est formé par trois sous unités E1, E2 et E3 et la protéine PDX1. Chez SACE, la sous unité E1 correspond à la fonction pyruvate

déshydrogénase, elle est codée par les gènes ScPDA1 YER178W (Steensma et al., 1990) et ScPDB1 YBR221C (Miran et al., 1993). E2 est une dihydrolipoamide acetyltransferase codée par ScLAT1 YNL071W (Niu et al., 1988), E3 une dihydrolipoamide dehydrogenase codée par ScLPD1 YFL018C (Ross et al., 1988) et PDX1 est codée par ScPDX1 YGR193C (Lawson et al., 1991). Cette voie permettant la synthèse d'acétyl-CoA à partir du pyruvate n'est pas décrite jusqu'ici chez YALI.

3.1.2 Les acétyl-CoA synthétases

Les hémiascomycètes ont la capacité de synthétiser l'acétyl-CoA avec des acétyl-CoA synthétases cytoplasmiques. Chez SACE il existe deux acétyl-CoA synthétase ScACS1 YAL054C et ScACS2 YLR153C. Ce sont des enzymes cytoplasmiques qui permettent de produire de l'acétyl-CoA à partir d'acétate. La délétion des gènes codant ces enzymes est létale sur toutes les sources de carbones chez SACE (van den Berg et al., 1996). Les ATP-citrate lyases ACL permettent également de produire de l'acétyl-CoA dans le cytoplasme mais à partir du citrate même en absence de source externe d'acétate. Les ACL sont principalement présents chez les algues, les plantes et les animaux. Des gènes codant des ACL sont présents dans tous les génomes fongiques analysés par Hynes *et al.* et sont absents chez tous les *Saccharomycotina* excepté chez YALI (Hynes and Murray). Effectivement une corrélation a été montré entre la présence d'ACL et le caractère oléagineux, c'est à dire la capacité à accumuler de grande quantité de lipides (Boulton and Ratledge, 1981).

Chez YALI un seul gène est homologue à ACS1 et ACS2, appelé ACS2 YALI0F05962g. Par contre YALI possèdent une ACL composée de deux sous unités codées par YlACL1 YALI0E34793g et YlACL2 YALI0D24431g (Zhou et al.).

3.1.3 La dégradation de la leucine

La dégradation de la leucine aboutit à la formation d'un acétyl-CoA dans la mitochondrie. La première étape est la transamination de la leucine par l'aminotransférase, codée par ScBAT1 YHR208W chez SACE, qui donne une cétoleucine (ou 4-methyl-2-oxopentanoate). La cétoleucine subit ensuite une déshydrogénation décarboxylante pour donner l'isovaleryl-CoA.

Chez les animaux, cette étape est réalisée par un complexe de la même famille que le complexe pyruvate déshydrogénase (PDA1, PDB1, LPD1, LAT1 et PDX1) impliqué dans le passage du pyruvate à l'acétyl CoA (voir paragraphe 3.1.1). Ce complexe déshydrogénase cétoacide-alpha à chaîne ramifiée (BCKDC) est composé de trois sous unités E1, E2 et E3. E1 composée de deux sous unités α et ß est une décarboxylase. E2 est une dihydrolipoyl transacylase et E3 est la même dihydrolipoamide dehydrogenase LPD1 que celle du complexe pyruvate déshydrogénase. Chez SACE les gènes codant ces enzymes ne sont pas identifiés, il est possible qu'ils n'existent

pas.

Chez les levures oléagineuse, l'isovaleryl CoA est ensuite déshydrogénée pour donner du ß-methylcrotonyl-CoA par l'isovaléryl-CoA déshydrogénase IVD (Vorapreeda et al.). Le ß-methylcrotonyl-CoA est ensuite carboxylé pour former du ß-methylglutaconyl-CoA. Les deux sous unités de la ß-methylcrotonyl-CoA carboxylase sont codées par MCC1 et MCC2 (Stadler et al., 2005) mais ne sont pas trouvées chez les organismes non oléagineux (Vorapreeda et al., 2012). Le ß-methylglutaconyl CoA est hydraté pour obtenir du 3-hydroxy-3méthylglutaryl-CoA puis sera clivé en acétoacétate et acétyl-CoA. Ces réactions sont réalisées chez les oléagineux par la methylglutaconyl-CoA hydratase MGH et l'hydroxymethylglutaryl-CoA lyase MVA. L'acétyl-CoA et l'acétoacétyl-CoA peuvent redonner du 3-hydroxy-3méthylglutaryl-CoA avec l'HMG-CoA synthase codée chez SACE par ScERG13 et le passage de l'acétyl-CoA à l'acétoacétyl-CoA se fait grâce à une acetoacetyl-CoA thiolase codée par ScERG10.

3.2 Transport de l'acétyl-CoA

L'acétyl-CoA est synthétisé dans la mitochondrie par la dégradation du pyruvate et de la leucine, dans le peroxysome par la ß-oxydation et l'ASC2 et dans le cytosol par les ACL. L'acétyl-CoA doit alors pouvoir passer d'un compartiment à l'autre afin de répondre aux besoins de la cellule. Pour cela il existe deux types de navettes : celle de la carnitine et celle du citrate malate.

3.2.1 La navette carnitine

Pour sortir des peroxysome l'acétyl-CoA peut utiliser les navettes carnitine. Chez SACE la carnitine entre dans le peroxysome et forme de l'acylcarnitine avec l'acétyl-CoA grâce à la carnitine transférase ScCat2p YML042W (Franken et al., 2008). L'acétylcarnitine peut alors être exportée du peroxysome vers les mitochondries grâce à la carnitine acétyl transférase ScYat1p YAR035W associée à la membrane externe de la mitochondrie et au transporteur de carnitine ScCrc1p YOR100C associé à la membrane interne (Figure 2.7).

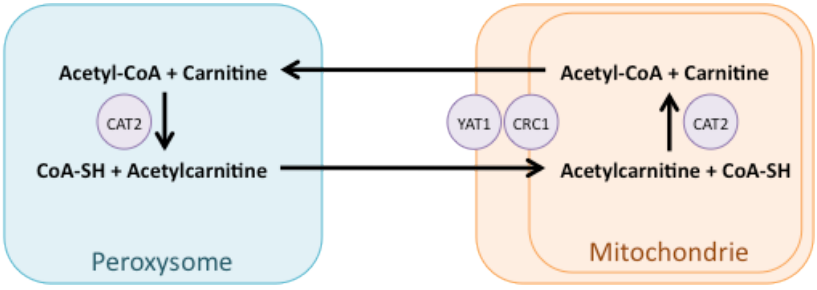


Figure 2.7 : Représentation schématique de la navette carnitine chez SACE adapté de Franken *et al.*. La carnitine transférase ScCat2p est à la fois dans le peroxysome et dans la mitochondrie, la carnitine acétyl transférase ScYat1p est sur la membrane externe de la mitochondrie et le transporteur de carnitine ScCrc1p sur la membrane interne de la mitochondrie (Franken et al., 2008).

La carnitine est synthétisée à partir de la triméthyl-lysine par une suite de réactions qui impliquent la trimethyl-lysine dioxygenase CNT, l'hydroxytrimethyl-lysine aldolase CNT2, l'aldehyde dehydrogenase cytosolique ALD et la butyrobetaine dioxygenase CNT. Des études ont montré que SACE ne semble pas avoir cette voie de biosynthèse de la carnitine. En effet, un mutant de SACE délété pour ScCIT2 YCR005C, codant la citrate synthase peroxysomale ne peut plus croître sur aucune sources de carbone si la carnitine n'est pas ajoutée au milieu (Swiegers et al., 2001; van Roermund et al., 1999). Cette incapacité à croitre s'explique par le fait qu'il n'existe dans le peroxysome que deux voies permettant la synthèse d'acétyl-CoA dont l'une est dépendante de la carnitine, puisque c'est la navette carnitine, et l'autre nécessite l'enzyme CIT2.

3.2.2 La navette citrate/malate

Pour sortir des mitochondries l'acétyl-CoA peut utiliser les navettes citrate/malate. La citrate synthase CIT1 YNR001C produit du citrate à partir de l'acétyl-CoA et l'oxaloacétate dans la mitochondrie. Le citrate peut alors sortir dans le cytosol pour redonner de l'acétyl-CoA à l'ATP-citrate lyase ACL chez les organismes oléagineux et/ou de l'oxaloacétate. Avec les malate déshydrogénases ScMDH1 mitochondriale et ScMDH2 cytosolique, codées par ScMDH1 YKL085W et ScMDH2 YOL126C chez SACE, l'oxaloacétate est converti en malate et inversement. La malate peut alors entrer dans la mitochondrie.

SACE possède trois gènes MDH permettant de convertir le malate en oxaloacétate. Chacun code pour des protéines avec des adressages différents: MDH1 YKL085W est adressé à la mitochondrie, MDH2 YOL126C au cytosol et MDH3 YDL078C au peroxysome (McAlister-Henn and Thompson, 1987; Minard and McAlister-Henn, 1991; Steffan and McAlister-Henn, 1992). ScMDH1 est impliquée dans la partie mitochondriale de la navette citrate/malate et dans le cycle des acides tricarboxyliques (TCA) (voir paragraphe 3.3.1); ScMDH2 dans la partie cytosolique de la navette citrate/malate, dans le cycle du glyoxylate, dans la navette redox (voir paragraphe 4.1) et dans la gluconéogenèse et ScMDH3 dans le cycle du glyoxylate et dans la navette redox.

YALI ne possède que deux gènes codant des malates déshydrogénases, YlMDH1 et YlMDH2. YlMDH1 code une protéine mitochondriale impliquée dans la navette citrate/malate et dans le cycle du TCA. YlMDH2 est alternativement épissé grâce à la variation de la taille du second intron (Figure 2.8). Ceci permet la production de deux formes de protéines : une cytosolique avec l'intron court, impliquée dans la navette citrate/malate et une peroxysomale avec l'intron long, impliquée dans le cycle du glyoxylate (Kabran et al., 2012).

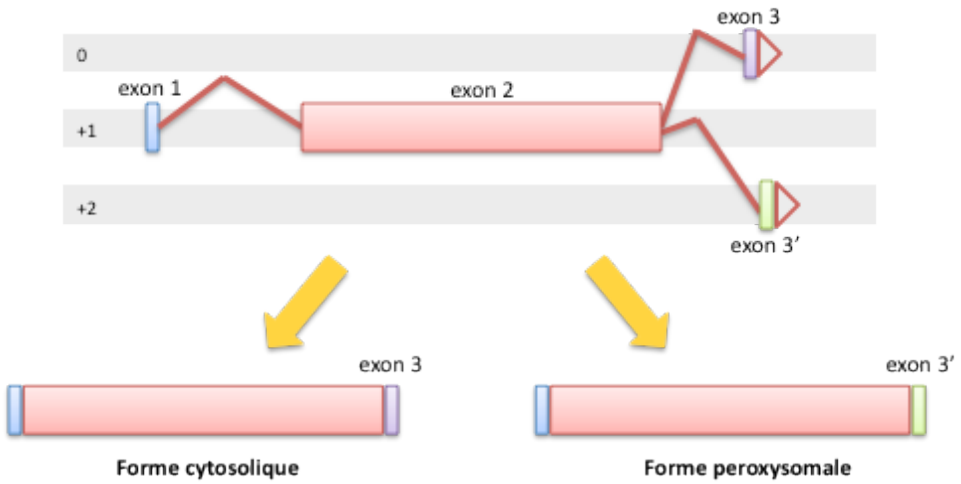


Figure 2.8 : Représentation schématique des transcrits alternatifs provenant du gène YlMDH2 d'après Kabran *et al.*. Les exons sont représentés avec des carrés de couleur et les introns sont symbolisés par des traits rouges. Les barres grises représentent les trois cadres de lecture 0, +1, +2 (Kabran et al., 2012).

3.3 L'utilisation de l'acétyl-CoA

3.3.1 Le cycle des acides tricarboxyliques (TCA)

L'acétyl-CoA permet d'assurer une partie des besoins énergétique de la cellule grâce au cycle du TCA qui se déroule dans les mitochondries. En effet le cycle du TCA utilise l'acétyl-CoA pour produire par une série de réactions des intermédiaires énergétiques qui servent à la production d'ATP.

La première étape de ce cycle est la condensation par la citrate synthase de l'oxaloacétate et de l'acétyl-CoA en citrate. Chez SACE il y a deux citrates synthases mitochondriales ScCIT1 YNR001C et ScCIT3 YPR001W (Jia et al., 1997), ScCIT2 YCR005C étant exclusivement peroxysomale (Lewin et al., 1990). ScCIT1 est une citrate synthase alors que ScCIT3 est à la fois une citrate synthase et une méthylcitrate synthase (Graybill et al., 2007).

Chez YALI il y également une méthylcitrate synthase mitochondriale YlCIT1 codée par YALI0E00638g. Par contre il n'y a qu'une seule citrate synthase, YlCIT2 codée par YALI0E02684g, mais avec une double localisation mitochondriale et peroxysomale (Kabran, 2011; Uchiyama et al., 1982).

Le citrate est ensuite isomérisé en isocitrate par des aconitases. Chez SACE il en existe deux ScACO1 YLR304C et ScACO2 YJL200C. Il est connu que ScACO1 est également impliqué dans la stabilité et l'organisation de l'ADN mitochondriale. Farooq *et al.* proposent que la délétion de ACO1 chez SACE active la voie de la régulation rétrograde, ce qui entraîne une augmentation de

la production de citrate par la régulation positive des gènes CIT1 ou CIT3. Cette augmentation du niveau de citrate conduirait donc à la surcharge en fer dans les mitochondries. Ce fer réagit ensuite avec du peroxyde d'hydrogène pour produire des radicaux hydroxyle, qui causent des dommages oxydatifs de l'ADN mitochondrial, et par conséquent son instabilité (Farooq et al., 2013). Chez YALI il y a également 2 aconitases, YlACO1 et YlACO2. Il a été montré en 2008 que YlACO1, homologue à ScACO1, code bien pour une aconitase fonctionnant dans le cycle de Krebs mitochondrial (Holz et al., 2009). Durant l 'étape suivante les isocitrates déshydrogénases oxydent et décarboxylent l'isocitrate pour obtenir de l'α-oxoglutarate. Chez SACE les isocitrates déshydrogénases NAD+ dépendantes IDH sont codées par ScIDH1 YNL037C et ScIDH2 YOR136W. Ces enzymes mitochondriales utilisent spécifiquement le cofacteur NAD+ et produisent du NADH. D'autres enzymes, les isocitrates déshydrogénases NADP+ dépendante IDP, permettent également le passage de l'isocitrate à l'α-oxoglutarate durant le cycle du TCA mais elles utilisent le cofacteur NADP+ et produisent du NADPH. Chez SACE il y a trois IDP mais seule ScIDP1 YDL066W est localisée dans la mitochondrie (Haselbeck and McAlister-Henn, 1991) alors que ScIDP2 YLR174W est cytosolique et ScIDP3 YNL009W est peroxysomale. ScIDP2 et ScIDP3 sont toutes deux impliquées dans la navette redox (Contreras-Shannon et al., 2005).

Chez YALI il n'y a qu'une seule IDP codée par YALI0F04095g. Elle partage 71 à 74% d'identité avec les IDP de SACE et d'après Yovkova *et al.* elle contient une séquence d'adressage aux mitochondries ce qui lui permet d'assurer le passage de l'isocitrate à l'α-oxoglutarate durant le cycle du TCA (Yovkova et al., 2013). Mais il a été observé au laboratoire que ce gène est capable de faire de l'épissage alternatif par rétention de l'intron en 5' UTR. Lorsqu'il y a rétention la traduction démarre plus tôt avec une méthionine en amont, incluant une séquence d'adressage aux mitochondries. Mais cette forme n'est présente que dans 5% des cas ce qui suggère que le passage de l'isocitrate à l'α-oxoglutarate se fait principalement avec des IDH NAD+ dépendantes et que YlIDP est majoritairement cytosolique (données non publiées de Mekouar M. et Neuvéglise C.).

Ensuite le complexe α-cétoglutarate déshydrogénase catalyse la formation du succinyl-CoA par décarboxylation oxydative de l'α-oxoglutarate. Cette réaction nécessite du coenzyme-A, du NAD et du FAD et produit du NADH. Le complexe enzymatique est composé de trois unités codées chez SACE par les gènes ScKGD1 YIL125W, ScKGD2 YDR148C et ScLPD1 YFL018C (Repetto and Tzagoloff, 1991) et chez YALI par les gènes fortement homologues à SACE YlKGD1 YALI0E33517g, YlKGD2 YALI0E16929g et YlLPD1 YALI0D20768g (Holz et al., 2011). Le succinyl-CoA est phosphorylé en succinate par la succinyl-CoA synthétase formée de deux sous-unités. La succinyl-CoA synthétase de SACE est codée par deux gènes : ScLSC1 YOR142W et ScLSC2 YGR244C mis en évidence par Przybyla-Zawislak, *et al.* (Przybyla-Zawislak et al., 1998). Ces auteurs montrent également que chez les levures cette réaction utilise strictement de l'ADP

et synthétise de l'ATP. Le succinate est ensuite converti en fumarate par oxydoréduction. Cette réaction est réalisée par la succinate déshydrogénase composée de quatre sous unités codées par ScSDH1, ScSDH2, ScSDH3 et ScSDH4. Chez YALI la succinate déshydrogénase est probablement aussi constituée de quatre sous unités mais seuls YlSDH1 et YlSDH2, codés par YALI0D11374g et YALI0D23397g, sont publiés bien que des homologues de SDH3 et SDH4 existent dans le génome: YALI0E29667g et YALI0A14784g (Yuzbashev et al., 2010). La fumarase, codée par ScFUM1, va donner, à partir du fumarate, du malate qui sera oxydé par la malate déshydrogénase pour donner de l'oxaloacétate. Otto *et al.* ont montré la présence du gène YlFUM1 YALI0C06776g chez YALI, codant pour une fumarase avec 74% d'identité avec SACE et un adressage à la mitochondrie (Otto et al., 2012). La malate déshydrogénase mitochondriale est codée par ScMDH1 chez SACE (Thompson et al., 1988) et par YlMDH1 chez YALI (Kabran et al., 2012). L'oxaloacétate pourra alors être réutilisé pour la première étape du cycle du TCA.

3.3.2 Le cycle du glyoxylate

Le cycle du glyoxylate permet de convertir les unités d'acétyl-CoA en composés à quatre carbones intermédiaires du cycle du TCA : malate, succinate et oxaloacétate (Figure 2.9). Le cycle du glyoxylate est alors un shunt du cycle du TCA qui se réalise dans le peroxysome et le cytosol. L'isocitrate formé au cours du cycle du TCA peut sortir de la mitochondrie pour être clivé par l'isocitrate lyase, codée par ScICL1 YER065C, en succinate et glyoxylate. Chez YALI cette enzyme est codée par le gène YlICL1 YALI0C16885g ayant bien un PTS1 (Kabran, 2011). Les étapes suivantes permettent de régénérer l'oxaloacétate à partir du glyoxylate par les actions successives de la malate synthase, codée par ScMLS1 YNL117W chez SACE et YlMLS2 YALI0D15708g chez YALI (Kabran, 2011) et de la malate déshydrogénase, codée par ScMDH3 YDL078C et par la forme peroxysomale de YlMDH2 (voir paragraphe 3.2.2).

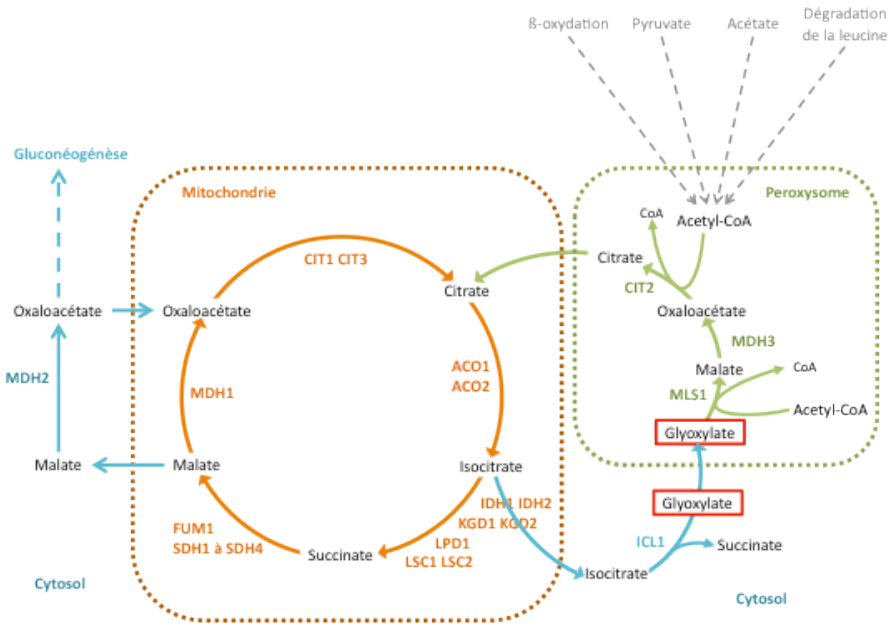


Figure 2.9 : Le cycle du glyoxylate chez SACE à partir de Koivistoinen *et al.* (Koivistoinen et al., 2013). Les enzymes mitochondriales sont en orange, les peroxysomales en vert et les cytosoliques en bleu. CIT2 : citrate synthase ; ACO1 : aconitase ; ICL1 : l'isocitrate lyase ; MLS1 : malate synthase ; MDH3 : malate déshydrogénase peroxysomale.

3.3.3 La synthèse des acides gras

L'acetyl-CoA est également essentiel pour la biosynthèse des acides gras car cette voie produit des acides gras par condensations successives de malonyl-CoA et d'acétyl-CoA réalisées par la fatty acyl-synthetase FAS. Ce cycle est décrit dans le paragraphe 5.2 : Le cycle d'élongation des acides gras.

4 Les métabolismes des cofacteurs : NADP et NADPH

Le NADH et le NADPH sont des coenzymes qui aident les enzymes à transférer les électrons pendant les réactions d'oxydo-réductions. Ce sont les sources principales d'électrons des cellules. Concernant le métabolisme lipidique, l'oxydation du NADPH est essentielle pour la réduction des AG par la ß-oxydation, pour les enzymes auxiliaires et pour la néo synthèse des AG par la FAS. Le NADP+ est réduit en NADPH par les IDP durant le cycle du TCA et lors de la conversion de la malate en pyruvate par l'enzyme malique YlMAE1 (Zhang et al.). D'après Ratledge *et al.* l'enzyme malique permettrait ainsi de contrôler le niveau d'accumulation des lipides chez les levures oléagineuses par l'importance de son rôle pour le pool de NADPH (Ratledge and Wynn, 2002). Le NAD+ est impliqué dans les oxydations lors de la ß-oxydation, de

la dégradation de la leucine, du cycle du TCA, du cycle du glyoxylate, de l'oxydation primaire des alcanes et de l'ω-oxydation. Le NADH est utilisé pour la production de G3P à partir de DHAP.

4.1 La navette redox

Chez SACE le NAD+ intervient lors de l'oxydation des alcools gras (acides gras dont le groupement carboxyle -COOH a été substitué par un groupement hydroxyle –OH), en aldéhydes gras (acides gras dont le groupement carboxyle -COOH a été substitué par un groupement aldéhyde -CH=O). Le NAD+ est également utilisé lors de la formation du 3-cétoacyl-CoA par l'activité 3-hydroxyacyl-CoA déshydrogénase de ScMFE2p. Dans les deux cas le NAD+ est réduit pour donner du NADH. De même, chez SACE, au cours de la ß-oxydation, le co-facteur NADPH est utilisé à deux étapes. La première est la réduction par la ScSps19p des dienoyl-CoA avec des doubles liaisons en position impaire. La seconde étape est la formation du 3-cétoacyl-CoA par l'action 3-hydroxyacyl-CoA déshydrogénase de la ScMFE2p. La navette redox permet alors de fournir du NADPH et du NAD+ dans les peroxysomes. Elle est composée de deux voies chez SACE. D'une part, le NAD+ est régénéré dans les peroxysomes grâce à l'échange malate/oxaloacétate. L'oxaloacétate entre dans le peroxysome où, en présence de NADH, il donne du malate et libère du NAD+. Le malate sort du peroxysome et régénère de l'oxaloacétate en présence du NAD+ cytosolique. L'oxaloacétate et/ou le malate sont fournis par le cycle du glyoxylate. Chez SACE cette voie est réalisée dans le cytosol par ScMDH2 et dans le peroxysome par ScMDH3. D'autre part, chez SACE seulement, le NADPH est fourni dans les peroxysomes grâce à l'échange céto-glutarate/isocitrate. Le 2-oxo-glutarate est réduit dans le cytosol en isocitrate par ScIdp1p en utilisant du NADPH. L'isocitrate est alors importé dans les peroxysomes où il est oxydé par ScIdp3p pour redevenir du 2-oxo-glutarate en libérant du NADPH (Henke et al., 1998).

Chez YALI la première voie pourrait être réalisée grâce aux deux formes de YlMDH2 générées par l'épissage alternatif. Par contre il n'y a pas d'IDP peroxysomale chez YALI donc la seconde voie n'est pas réalisable. Kabran *et al.* ont d'ailleurs observé que la souche exprimant seulement la MDH cytosolique a le même phénotype que la souche sauvage (Kabran et al., 2012). Ainsi la MDH peroxysomale n'est pas essentielle pour la réoxydation du NADH dans les peroxysomes. Ils suggèrent donc qu'il existe une voie alternative à cette navette redox dans les peroxysomes qui pourrait utiliser la lactate déshydrogénase ou la glycerol-3-phosphate déshydrogénase, présents dans les mammifères n'ayant pas de MDH.

4.2 La navette citrate-oxoglutarate dans la mitochondrie

En 2010, Castegna *et al.* ont découvert une nouvelle navette dans la mitochondrie de SACE (Figure 2.10) (Castegna et al.). Dans la mitochondrie, l'oxoglutarate avec du NADPH et du CO2 donne du citrate en libérant du NADP+. Le citrate peut alors sortir dans le cytosol grâce au transporteur mitochondrial YHM2p YMR241W. Le citrate redonne alors de l'oxoglutarate en utilisant du NADP+ et en libérant du NADPH + CO2. Le NADP+ libéré dans la mitochondrie peut alors être utilisé lors du cycle de Krebs directement relié au cycle du glyoxylate.

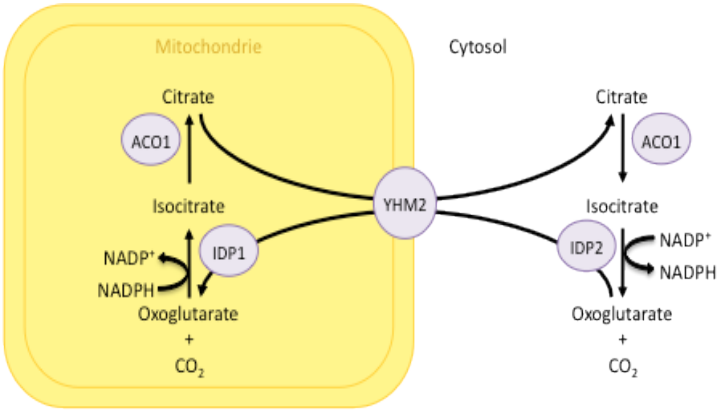


Figure 2.10 : Réactions impliquées dans la navette citrate-oxoglutarate de la mitochondrie au cytosol chez SACE (Castegna et al., 2010). ScYHM2 est sur la membrane de la mitochondrie. ScACO1 est mitochondriale et cytosolique. ScIDP1 est mitochondriale et ScIDP2 cytosolique.

4.3 La navette malate/oxaloacétate

La malate et l'oxaloacétate peuvent entrer et sortir de la mitochondrie vers le cytosol. Le passage de la malate à l'oxaloacétate par la malate déshydrogénase MDH donne un NADH et le passage de l'oxaloacétate à la malate génère du NAD+. La déshydrogénation de la malate dans la mitochondrie est une réaction qui fait partie du cycle du TCA. L'oxaloacétate dans le cytosol pourra entrer dans la navette redox ou la gluconéogenèse ou le cycle du glyoxylate. Mais il peut aussi être converti de façon réversible en pyruvate par la pyruvate décarboxylase PDC1, la réaction inverse étant réalisée par la pyruvate carboxylase PYC1. Le pyruvate permet à la cellule de produire de l'acétyl-CoA. Chez SACE et YALI, les MDH impliquées sont la MDH2 dans le cytosol et la MDH1 dans la mitochondrie.

5 SYNTHÈSE *DE NOVO* DES ACIDES GRAS

5.1 Le déclenchement de l'accumulation

Chez les levures oléagineuses, la présence d'une quantité suffisante d'azote dans le milieu diminue l'activité de l'AMP désaminase et donc augmente la présence d'AMP. L'AMP active les isocitrates déshydrogénases IDH1 et IDH2 qui oxydent l'isocitrate en α-cétoglutarate au cours du cycle du TCA (Ratledge and Wynn, 2002).

Inversement, un faible taux d'azote dans le milieu diminue la présence d'AMP par l'activation de l'AMP désaminase et donc celle des IDH. L'isocitrate produit lors du cycle du TCA ne pourra pas être oxydé pour continuer le cycle et sera alors importé dans la mitochondrie par la navette malate/citrate (voir paragraphe 3.2.2). L'acétyl-CoA produit par cette navette sera ainsi utilisé pour la synthèse *de novo* des acides gras.

5.2 Le cycle de la néo synthèse des acides gras

La synthèse des acides gras permet l'obtention d'acides gras saturés à longue chaine carbonée à partir de l'acétyl-CoA, du malonyl CoA et du NADPH (voir (Tehlivets et al., 2007). Le cycle permettant cette synthèse est réalisé dans la mitochondrie et le cytosol. La première étape est le passage de l'acetyl-CoA au malonyl-CoA. Il est toujours assuré par l'acétyl-CoA carboxylase codée par ScACC1 YNR016C chez SACE et YlACC1 YALI0C11407g chez YALI. YlACC1 est une enzyme clé de la synthèse des lipides car elle détourne les flux vers cette synthèse et sa surexpression augmente fortement la production de lipides chez YALI (Tai and Stephanopoulos, 2013).

Le cycle de synthèse est réalisé par le complexe enzymatique de la synthase d'acides gras FAS. Il existe deux types de complexe FAS : la FAS I cytosolique et la FAS II mitochondriale.

La FAS I est un homodimère multifonctionnel codé par deux gènes ScFAS1 et ScFAS2. Chez les levures la FAS I est capable de synthétiser des acides gras à 20 carbones *in vitro* mais elle n'est activée qu'en absence d'acides gras exogènes. La FAS II est constituée d'enzymes chacune étant monofonctionnelle codées chez SACE par ScMCT1, ScCEM1, ScOAR1, ScHTD2 et ScETR1. Les acides gras sont synthétisés sous forme d'acyl-CoA par la répétition des réactions suivantes : condensation, réduction, déshydratation et réduction qui sont toutes assurées par la FAS II dans la mitochondrie.

En 2013, Kursu *et al.* ont montré que FAS II est également impliquée dans les mécanismes d'épissage de l'ARNm, de la traduction mitochondriale et de l'assemblage des complexes respiratoires. Ils proposent alors un modèle où la FAS II serait comme un capteur mitochondrial pour la disponibilité en acétyl-CoA et un coordinateur de l'expression des gènes nucléaires et

mitochondriaux en adaptant le compartiment mitochondrial aux changements d'état métabolique de la cellule (Kursu et al., 2013). Chez YALI le système FASII mitochondriale n'a jamais été décrit.

Dans le cytosol la FASI permet la condensation et la première réduction avec l'enzyme FAS2 et la déshydratation puis la réduction avec FAS1. Chaque cycle ajoute 2 carbones à l'acyl-CoA mais l'élongation s'arrête en général au bout de 7 cycles. Les groupes acyl qui ont alors 16 carbones peuvent subir différentes modifications notamment sous l'action des élongases et des désaturases. Ces modifications ont lieu dans les peroxysomes et permettent la formation des acides gras saturés ou insaturés à très longues chaînes (VLCFA) c'est à dire avec 20 carbones ou plus. Chez YALI il existe deux élongases ELO1, qui permet de passer d 'un C16 à un C18, et ELO2 qui permet les élongations suivantes jusqu'au C24. Elles sont respectivement codées par YlELO1 et YlELO2. Les désaturases sont aussi de deux sortes : la désaturase Δ9 YlOLE1 et la désaturase Δ12 YlFAD2 (Anamnart et al., 1997; Beopoulos et al., 2014; Morin et al., 2011).

YALI est dépourvu d'acide gras ω-3 désaturase contrairement à *Pichia Pastoris* (Zhang et al., 2008) ou *Candida albicans* par exemple . Une nouvelle sorte d'acide gras ω-3 désaturase Δ17 a récemment été identifiée chez trois espèces d'oomycetes. Les auteurs ont caractérisé ces 3 enzymes en les exprimant chez YALI afin de montrer leur forte activité. Ces Δ17 désaturases peuvent également convertir des acides gras ω-6 en acides gras ω-3.

6 ACCUMULATION SOUS FORME DE LIPIDES NEUTRES

La plupart des eucaryotes mettent en réserve des lipides ce qui leur permet d'avoir d'importantes réserves énergétiques. Les lipides sont mis en réserve sous forme de triglycérides et d'esters de stérols plutôt que sous forme de glucides. En effet leur catabolisme par oxydation s'accompagne d'une production d'énergie 2 fois plus grande et leur stockage se fait sous une forme physique compacte et sans eau. Ces lipides neutres sont stockés dans les corps lipidiques chez les plantes, certaines levures comme SACE et les levures oléagineuses comme YALI. L'accumulation des lipides neutres est un métabolisme connu et bien renseigné en terme de gènes chez SACE (Czabany et al., 2007) et chez YALI (Athenstaedt, 2011; Beopoulos et al., 2012; Beopoulos et al., 2008; Dulermo and Nicaud). Ce chapitre s'appuie donc sur ces publications.

6.1 Synthèse des triglycérides

Les triglycérides sont synthétisés par la voie de Kennedy à partir du glycérol-3-phosphate G3P (Figure 2.11).

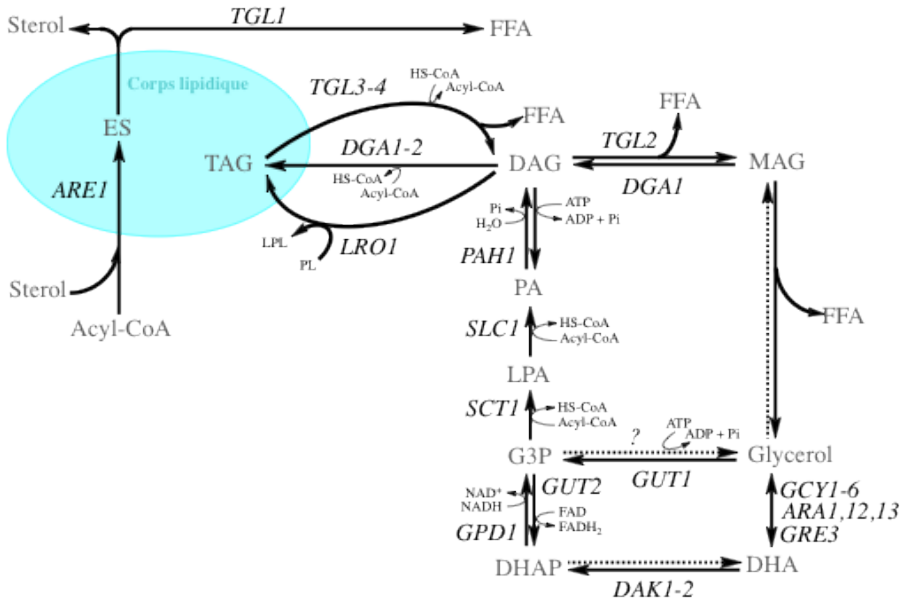


Figure 2.11 : Représentation schématique de la synthèse des TAG et de la synthèse de ES chez YALI d'après Dulermo *et al.*. Les composées sont représentés en gris : ES : ester des stérol ; FFA : acide gras libre ; MAG, DAG, TAG : mono-, di-, tri-glycéride ; PA : acide phosphatidique, LPA : acide lysophosphatidique ; G3P : glycérol-3-phosphate ; DHAP : déshydroxyacétone phosphate ; DHA : dihydroxyacétone. Les enzymes sont représentées par leurs gènes en noir italique : ARE1 : acyl-CoA stérol acyltransférases ; TGL1, TGL2, TGL3, TGL4 : lipases intracellulaires ; DGA1, DGA2, LRO1 : DAG-acyltransférases ; PAH1 : PA-déphosphorylase ; SLC1 : LPA-acyltransférase ; SCT1 : G3P-acyltransférase ; GPD1 : G3P-déshydrogenase NAD+ dépendante ; DAK1, DAK2 : glycérones kinases ; GCY1 à GCY6, ARA1, ARA12, ARA13, GRE3 : glycérol déshydrogénase. (Dulermo and Nicaud, 2011)

Le G3P peut être synthétisé directement à partir du glycérol par la glycérol kinase GUT1 codée par ScGUT1 YHL032C et YlGUT1 YALI0F00484g. Le glycérol peut aussi être converti en dihydroxyacétone DHA par la glycérol déshydrogénase codée chez SACE par ScGCY1 YOR120W, ScYPR1 YDR368W, ScARA1 YBR149W et ScGRE3 YHR104W. Chez YALI il y a 6 gènes GCY (YlGCY1 YALI0D04092g, YlGCY12 YALI0B07117g, YlGCY13 YALI0A15906g, YlGCY14 YALI0E18348g, YlGCY15 YALI0F18590g, YlGCY16 YALI0B21780g), trois gènes ARA (YlARA1 YALI0F06974g, YlARA12 YALI0C13508g, YlARA13 YALI0C09119g) et YlGRE3 YALI0D07634g. Le DHA est ensuite phosphorylé en déshydroxyacétone phosphate DHAP par les dihydroxyacétones kinases DAK1 (YML070W, YALI0E20691g) et DAK2 (YFL053W, YALI0F09273g). Le DHAP est alors converti en G3P par la G3P-déshydrogénase YlGUT2 YALI0B13970g et ScGUT2 YIL155C (la réaction inverse est réalisée par la G3P-déshydrogenase NAD+ dépendante codée par ScGPD1 YDL022W et ScGPD2 YOL059W chez SACE et par un seul gène YlGPD1 YALI0B02948g chez YALI). Le G3P peut alors entrer dans la voie de synthèse des TAG appelée la voie de Kennedy. Le G3P subit une double acylation par les actions successives de la G3P-acyltransférase SCT1 (YBL011W, YALI0C00209g) et la LPA-acyltransférase SLC1

(YDL052C, YALI0E18964g) pour former de l'acide phosphatidique PA. La PA est alors déphosphorylé en DAG par la PA-déphosphorylase PAH1 (YMR165C, YALI0D27016g). Les triglycérides TAG sont ensuite synthétisés directement à partir des diglycérides DAG par acylation. Si l'acylation utilise l'acyl-CoA comme donneur d'acyle, elle est réalisée par les DAG-acyltransférases ScDGA1 YOR245C chez SACE et YlDGA1 YALI0E32769g et YlDGA2 YALI0D07986g chez YALI. Si elle utilise un glycérophospholipide comme un donneur d'acyle elle est réalisée par la DAG-acyltransférase LRO1 (YNR008W, YALI0E16797g). Cette acylation est réversible grâce à deux lipases intracellulaires TGL3 (YMR313C, YALI0D17534g) et TGL4 (YKR089C, YALI0F10010g) (Dulermo et al., 2013).

6.2 Synthèse d'esters de stérols

La synthèse des esters de stérols ES chez SACE est réalisée par la condensation de l'acyl-CoA avec du stérol par deux isoenzymes acyl-CoA stérol acyltransférases codées par ScARE1 YCR048W et ScARE2 YNR019W. Chez YALI il y a une seule acyl-CoA stérol acyltransférases codée par YlARE1 YALI0F06578g (Figure 2.11). Les ES peuvent redonner des stérols et des acides gras libres par la lipase intracellulaire ScTGL1 YKL140W chez SACE.

6.3 Les corps lipidiques

Les TAG et les SE ne peuvent pas s'intégrer dans la bicouche lipidique. Ils sont alors stockés dans les corps lipidiques constitués d'une monocouche de phospholipides contenant quelques protéines entourant le noyau formé par les lipides neutres. Chez SACE les corps lipidiques contiennent autant de TAG que de SE alors que chez YALI il n'y a que très peu de SE (5% de SE pour 95% de TAG) (Athenstaedt et al., 2006; Leber et al., 1994). Mais les corps lipidiques ne sont pas de simples particules de stockage fournissant des lipides en cas de manque dans le milieu. En effet de nombreuses protéines à fonction connues impliquées dans le métabolisme lipidique ont été retrouvées dans les corps lipidique. Par exemple chez SACE Athenstaedt *et al.* ont analysé les protéines majeures se retrouvant dans les corps lipidiques et ils y ont trouvé les acyl-CoA synthétases ScFAA1 et ScFAA4 et le transporteur d'acides gras ScFAT1 impliqués dans l'import des LCFA dans la cellule (Athenstaedt et al., 1999) (voir paragraphe 2.1). Actuellement nous savons que chez SACE des protéines contribuant à l'activation des acides gras, à la synthèse et la dégradation des TAG, au métabolisme des stérols et à la synthèse de l'acide phosphatidique sont présentes dans les corps lipidiques et que la suppression de certaines protéines entraîne des changements dans la quantité de lipides stockés (Athenstaedt et al., 2006).

Chez YALI on sait que la composition lipidique et protéique varie en fonction des milieux de croissance et plus particulièrement en fonction de la source de carbone. Le changement de glucose à acide oléique par exemple induit à la fois un grossissement des corps lipidiques et un changement dans la composition lipidique et protéique (Athenstaedt et al., 2006).

III. RÉSULTATS

1 LES 4 NOUVEAUX GÉNOMES DU CLADE…ET CELUI DE OLHI

Pour étudier le métabolisme lipidique des espèces du clade *Yarrowia*, 5 génomes complets ont été entièrement séquencés avant mon arrivée et annotés au cours de ma thèse au sein de l'équipe. Au cours de la première année de thèse nous considérions que OLHI faisait partie du clade *Yarrowia* mais au vu de sa divergence en terme de génome nous la considérons maintenant comme une espèce extérieure au clade. Nous avons subdivisé le clade en deux groupes : les espèces les plus proches de YALI que nous appelons les *sensu stricto*, comprenant notamment YAGA capable de croiser avec YALI (Knutsen et al., 2007) et les espèces plus éloignées que nous appelons les *sensu lato*. Ainsi, parmi ces 5 génomes, deux proviennent des espèces *sensu stricto* YAGA et YAYA, deux des *sensu lato* YAPH et YAAL et un de l'espèce externe au clade OLHI. L'annotation a été réalisée dans un premier temps par transfert d'annotation en utilisant Amadea Biopack (ISoft, France) puis par curation manuelle facilitée par les données de RNASeq par plusieurs membres de l'équipe.

Les génomes ont une bonne couverture, peu de larges scaffolds (1 à 2 par chromosome) et pour certaines espèces très peu de gaps (Tableau 2.1). Le contenu en GC varie autour des 49 % pour les *sensu stricto* et YAAL. Il est de 43 % pour YAPH et 41% pour OLHI. L'annotation a permis de détecter de 5916 à 6936 CDS par génome sauf pour OLHI qui a environ 1000 gènes de moins. La taille des génomes est aussi très variable, environ 20 Mb pour les espèces les plus proches de YALI et seulement 10.7 Mb pour OLHI. Ceci suggère une expansion de la taille des génomes au sein du clade *Yarrowia*.

Tableau 2.1 : Présentation des génomes utilisés pour l'étude.

		Couverture	Gaps	Scaffolds (>5 kb)	Contenu en GC	Taille du génome (Mb)	CDS (dont pseudo)	
Y. lipolytica	Sanger	10 X	11	6	49 %	20.5	6642 (141)	
C. galli CBS 9722	454 single read + paired-end 8kb	23.9 X	177	6	49 %	22.8	6936 (153)	*SENSU STRICTO*
C. yakushimensis CBS 10253	454 single read + paired-end 8kb + HiSeq 50 bp	28.2 X	33	8	48 %	18.8	6283 (70)	
C. phangngensis CBS 10407	454 single read + paired-end 8kb	32.6 X	111	9	43 %	16.2	5916 (148)	*SENSU LATO*
C. alimentaria CBS 10151	454 single read + paired-end 8kb + HiSeq 50 bp	16.7 X	116	11	49 %	19.8	6209 (128)	
C. hispaniensis Y-5580T	454 single read + paired-end 8kb + HiSeq 50 bp	39.5 X	17	6	41 %	10.7	5338 (64)	OUT GROUP

Ce phénomène d'expansion a été étudié au sein de l'équipe. Il ne peut pas s'expliquer par la différence du nombre de gènes entre les espèces car OLHI n'a que 1300 gènes de moins que les espèces qui ont un génome 2 fois plus grand. Cette expansion n'est pas due à la présence d'éléments transposables puisqu'ils ne représentent que 1.5% du génome de *Y. lipolytica*. Et elle n'est pas due à une duplication complète du génome ancestral car nous n'observons pas la présence d'ohnologues, ni de blocs de duplication et il y a peu de pseudogènes. Par contre, si l'on représente la taille moyenne des intergènes en fonction de la taille des génomes, comme présenté sur la figure 2.12, nous constatons que OLHI se situe dans le même ensemble que les autres hémiascomycètes telles que *S. cerevisiae, L. kluyveri* ... Par contre les 5 espèces du clade *Yarrowia* ont subi une expansion de la taille de leur génome qui est directement liée à la taille moyenne de leurs intergènes.

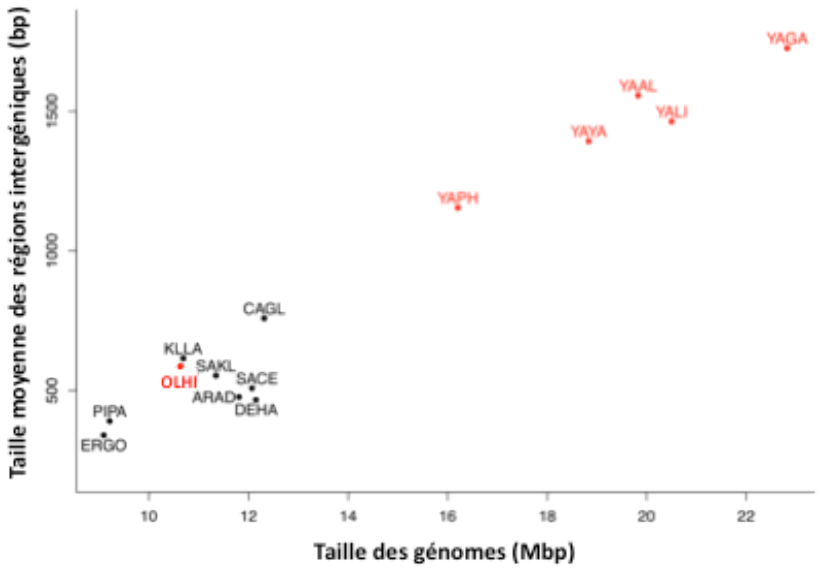


Figure 2.12 : Corrélation entre la taille des génomes et la taille moyenne des régions intergéniques. En rouge sont représentées les espèces du clade *Yarrowia* et en noir les autres espèces hémiascomycètes.

L'accès aux génomes des 6 espèces YALI, YAGA, YAYA, YAPH, YAAL et OHLI a permis de réaliser une phylogénie du clade (Figure 2.13) basée sur l'alignement de 912 gènes codant des protéines, présents en en copie unique dans chaque génome (398959 résidus) (voir encadré page 73). Cet alignement est également utilisé pour calculer les pourcentages de similarité et d'identité entre chaque couple d'espèces. L'emplacement de YAGA et de YAYA n'est pas en accord avec la phylogénie réalisée à partir de sept marqueurs dont EF1α proposée dans l'article publié « Comparative Physiology of Oleaginous Species from the *Yarrowia* Clade » (Figure 2.13). Ceci s'explique par le fait que les espèces *sensu stricto*, qui partagent en moyenne 90% de similarité, sont extrêmement proches et donc difficiles à placer sur un arbre. Les espèces *sensu lato* du clade *Yarrowia* sont séparées, entre elles et avec les *sensu stricto,* par plusieurs millions d'années. Nous avons déterminé pour chaque couple d'espèces la moyenne des pourcentages de similarité des paires de RBH (reciprocal best hit) synténiques déterminés par SynChro (Drillon et al., 2014). Le pourcentage de similarité entre les espèces *sensu lato* YAPH et YAAL est de 79% et il est de 82% entre YALI est YAPH. OLHI par contre partage en moyenne 67% de similarité avec toutes les espèces du clade (Tableau 2.2).

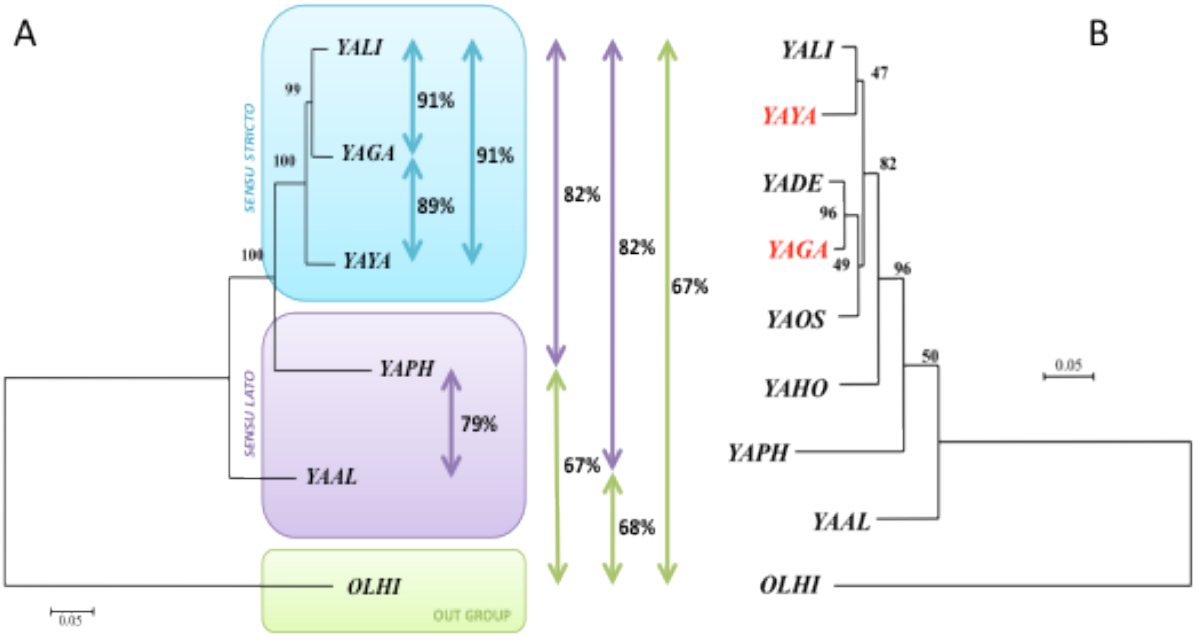


Figure 2.13 : A) Arbre phylogénétique du clade *Yarrowia*, construit par concaténation de 912 familles de gènes, 398959 résidus, 100 réplicats. Les pourcentages indiquent la similarité entre les espèces, calculée à partir des orthologues synténiques. B) Arbre phylogénétique des 9 espèces du clade *Yarrowia*, présenté dans le Chapitre 1, construit par concaténation de sept marqueurs (2207 acides aminés), 100 réplicats. Les espèces n'ayant pas le même emplacement dans les deux arbres sont notées en rouge.

Tableau 2.2 : Pourcentage de similarité entre les génome des espèces du clade *Yarrowia* et OLHI. Les pourcentages sont calculés à partir des orthologues synténiques de chaque paire d'espèces.

	YAGA	YAYA	YAPH	YAAL	OLHI
YALI	91,5	90,6	82,4	82,2	67,5
YAGA		89,5	82,1	81,9	67,5
YAYA			81,8	81,6	67,5
YAPH				79,5	67,5
YAAL					67,7

Les dot-plots des génomes complets entre les espèces du clade montrent de nombreux réarrangements inter- et intrachromosomiques. En alignant les génomes des espèces *sensu stricto* contre celui de YALI nous pouvons observer que l'ordre des gènes est globalement bien conservé le long des chromosomes bien que beaucoup de bras subissent déjà des réarrangements intrachromosomiques (Figure 2.14). Lorsque l'on compare YALI aux espèces *sensu lato* nous observons beaucoup de mésosynténie, particulièrement entre YALI et YAAL qui sont les espèces les plus éloignées du clade. Entre YALI et OLHI la synténie n'est plus conservée.
L'étude des blocs de synténie confirme ces nombreux réarrangements. En effet, même si une forte proportion de gènes orthologues se retrouve dans des blocs de synténie, le nombre de blocs est très important, de 91 pour le couple YAYA-YALI à 794 pour YAGA-YAAL (Tableau 2.3). Le nombre d'orthologues synténiques varie de 5996 à 5807 pour les paires d'espèces *sensu stricto* et de 5601 à 5132 pour les paires impliquant des espèces *sensu lato*. OLHI ne possède que 4586 gènes en synténie avec YAAL et au maximum 4671 avec YAYA. YAYA est l'espèce qui possède le plus de gènes impliqués dans des blocs de synténie et a en moyenne le plus petit nombre de blocs. Ceci suggère que YAYA possède l'organisation du génome la plus proche de celle de l'ancêtre commun aux 6 espèces.

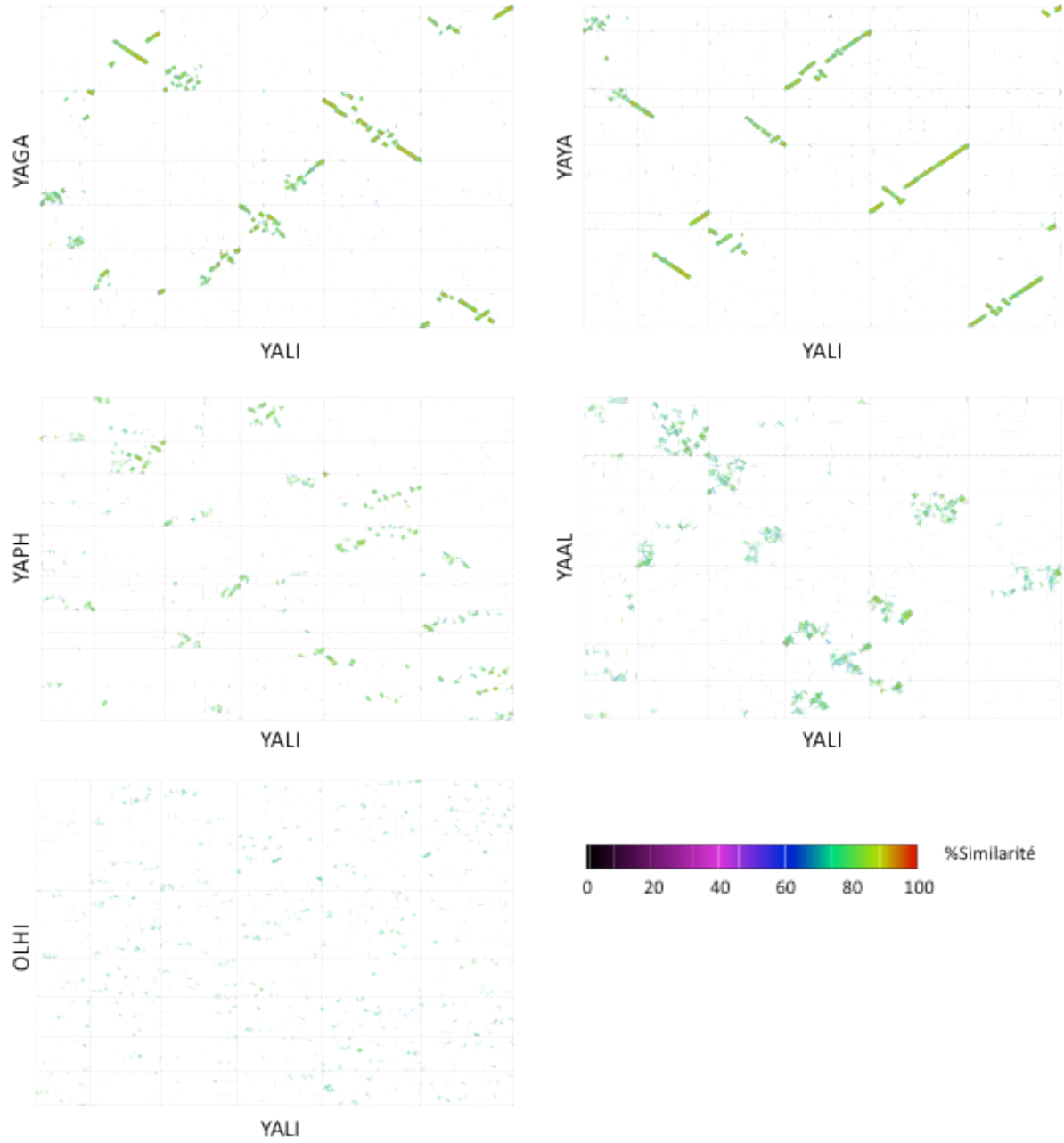


Figure 2.14 : Dot-plots des génomes complets de YAGA, YAYA, YAPH, YAAL et OLHI contre YALI.

	YALI	YAGA	YAYA	YAPH	YAAL	OLHI
YALI		198	91	549	777	888
YAGA	5996		218	588	794	887
YAYA	5819	5807		558	773	891
YAPH	5237	5196	5271		771	912
YAAL	5601	5588	5537	5132		786
OLHI	4613	4595	4671	4685	4586	

Tableau 2.3 : Pour chaque couple d'espèces du clade *Yarrowia* : en bas à gauche, le nombre de RBH synténiques est donné en noir; en haut à droite, le nombre de blocs de synténie, en bleu. Les espèces *sensu stricto* sont en vert.

Reconstructions Phylogénétiques

Pour reconstruire la phylogénie du clade, un ensemble de séquences de 912 protéines à été choisi selon les critères suivants : les gènes sont en singleton dans tous les génomes disponibles dans Génolevures (sauf chez PISO), l'alignement des séquences protéiques, nettoyé avec Gblocks (Castresana, 2000), représente au moins 70% de l'alignement initial et l'alignement résultant contient au moins 100 acides aminés. Les 912 alignements sont ensuite concaténés ce qui conduit à un alignement de 398959 résidus. Les arbres phylogénétiques ont été construits par maximum de vraisemblance, avec PhyML (Guindon et al., 2010) et un modèle de substitution JTT corrigé pour l'hétérogénéité entre les sites par une distribution Γ-law, avec quatre catégories de taux d'évolution. La proportion des sites invariables et l'α-paramètre de la distribution de la Γ-law ont été optimisés en fonction des données. Une valeur de bootstrap a été calculée à partir de 100 répétitions.
Pour les phylogénies de gènes présentées dans cette thèse, les arbres ont été construits avec PhyML selon les mêmes paramètres utilisés pour la phylogénie des espèces.

2 STRATÉGIES

2.1 Reconstruction du métabolisme lipidique

Certaines étapes du métabolisme lipidique sont très bien caractérisées d'un point de vue fonctionnel et protéique mais les gènes associés à ces fonctions ne sont pas toujours connus et des listes de noms systématiques de gènes associés aux fonctions étudiées ne sont que trop peu souvent fournies. Pour reconstituer le métabolisme lipidique de YALI nous avons décidé de nous intéresser aux voies impliquées dans le métabolisme des alcanes et des acides dicarboxyliques (DCA), dans la synthèse, l'élongation, le stockage et la dégradation des AG et le cycle de l'acétyl-CoA. Pour couvrir la plus grande partie du cycle de l'acetyl CoA nous avons également inclus le cycle du TCA, la synthèse de la carnitine, la dégradation de la leucine et du pyruvate.
La stratégie était de combiner les données de bibliographie, les données disponibles sur KEGG et les annotations fonctionnelles des gènes de YALI et de SACE pour d'une part reconstituer le métabolisme et d'autre part pour déterminer les gènes impliqués dans chacune des réactions nécessaires. Pour les parties du métabolisme pour lesquelles aucune étude génétique ou protéique n'avait été réalisée, la stratégie consistait à :
- Réaliser des blast à partir de protéines dont l'activité enzymatique est connue chez d'autres levures ou champignons pour identifier des gènes candidats ;

- Regarder les familles Génolevures incluant les gènes impliqués dans le métabolisme lipidique et dont on ne connaît pas les homologues chez YALI pour identifier des gènes candidats ;
- Déterminer l'adressage des protéines ;
- Regarder l'expression des gènes dans les données de transcriptomique ;
- Rechercher la présence de gènes candidats dans les autres espèces du clade.

Les gènes du métabolisme lipidique de YAGA, YAYA, YAPH, YAAL et YAHI sont retrouvés par homologie en utilisant les gènes de YALI. Les orthologues dans chaque espèce sont identifiés par une approche par blastp réciproque en deux étapes (Figure 2.15). Dans la première étape, les gènes de YALI sont utilisés comme référence pour réaliser un blastp contre les 5 autres espèces avec un cut-off E-value de $1.e^{-10}$ (Altschul et al., 1990). Ensuite un blastp réciproque ne considérant que le meilleur hit permet d'exclure les gènes correspondant à d'autres gènes de YALI.

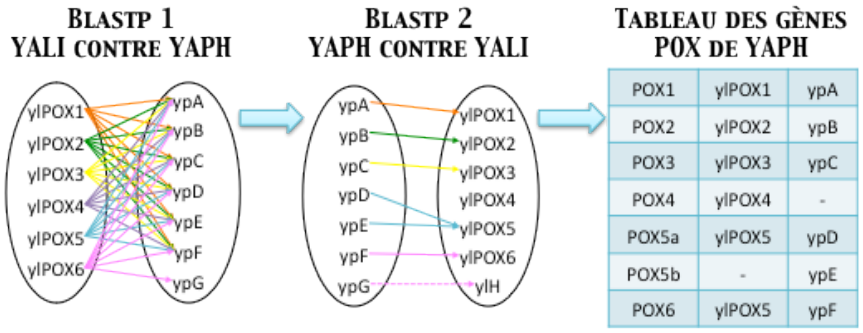


Figure 2.15 : Stratégie utilisée pour retrouver les gènes du métabolisme lipidique dans les cinq génomes en utilisant les gènes de YALI. Cette figure prend pour exemple les gènes POX dans les génomes de YALI et YAPH. Le Blastp 1 est réalisé avec un cutoff E-value de $1.e^{-10}$ et le Blastp 2 ne prend que le meilleur hit.

Au vu de la complexité des familles de gènes impliquées dans le métabolisme lipidique dans le clade *Yarrowia*, une étape de curation manuelle après ce blast réciproque a été nécessaire pour attribuer correctement les bons orthologues dans les familles de gènes.

2.2 Définition des familles de gènes

Génolevures fournit un classement en familles des protéines des génomes disponibles sur le site, dont YALI. Ces familles sont définies selon les comparaisons des séquences entières des protéines et selon les regroupements consensus algorithmiques, selon la méthode définie par Nikolski et Sherman (Nikolski and Sherman, 2007). Cette approche n'était donc pas assez précise pour les études sur les familles que nous avons menées au cours de ce projet de thèse car une famille est définie à partir de génomes très différents et une famille peut inclure à la fois des gènes du métabolisme lipidique et des gènes d'un autre métabolisme. Nous avons alors défini de

nouvelles familles basées sur celles de Génolevures mais redéfinies plus finement par rapport aux génomes du clade, donc souvent découpées en plusieurs familles, et qui ne contiennent a priori que des gènes impliqués dans le métabolisme lipidique.

Par exemple, la famille GL3C0072 de Génolevures regroupe 39 membres dont 3 chez SACE et 14 chez YALI. Chez SACE ces protéines sont les acetyl-CoA synthetases ACS1 et ACS2 et l'oxalyl-CoA synthétase FAT2. Chez YALI cette famille regroupe les 10 coumarates-CoA ligases YlCOU1 à YlCOU10, YlFAT3, YlFAT4 et YlFAT4, annotées comme similaires à ScFAT2, et YlACS2. Sur la représentation des relations au sein de la famille, fournie par Génolevures, nous observons 4 groupes distincts dont un comprenant les deux ACS de SACE et YlACS2 ; le second regroupe ScFAT2, YlFAT4 et YlFAT2 ; le troisième, relativement proche du second groupe, est composé de la seule protéine YlFAT3 et le dernier groupe comprend exclusivement les 10 coumarates-CoA ligases de YALI. Nous avons alors construit l'arbre phylogénétique des protéines de SACE et de YALI de la famille GL3C0072 (Figure 2.16). Sur cet arbre nous constatons qu'effectivement les coumarates-CoA ligases constituent un groupe monophylétique éloigné des FAT et des ACS. Nous estimons donc que les coumarates-CoA ligases constituent une famille différente. Les ACS sont très éloignées des FAT et des coumarates-CoA ligases, nous avons donc décidé de retirer YlACS2 des familles et d'en faire un singleton. Les FAT ne constituent pas un groupe monophylétique mais sont proches entre elles et ont une fonction similaire, nous les avons donc groupées dans une même famille.

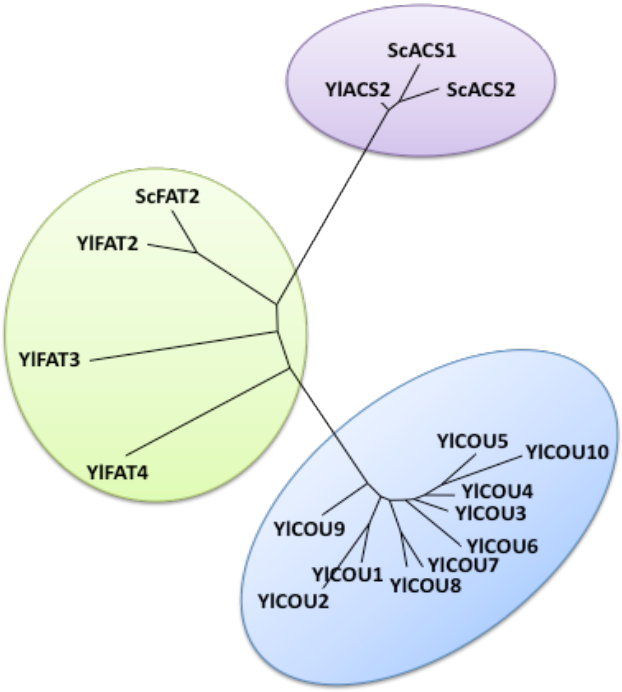


Figure 2.16 : Arbre phylogénique des coumarates-CoA ligases de SACE et YALI de la famille Génolevures GL3C0072, 931 résidus.

3 RECONSTRUCTION DU MÉTABOLISME LIPIDIQUE

La reconstruction du métabolisme lipidique de YALI a donné lieu à un schéma présentant toutes les voies métaboliques impliquées, donné en Annexe1. Les gènes codant les enzymes nécessaires à ce métabolisme sont répertoriés dans le tableau en Annexe 2.

Chez YALI nous recensons 204 gènes impliqués dans le métabolisme des lipides (Tableau 2.4). Parmi ces 204 gènes 136 sont répartis en 37 familles de gènes. Ces familles de gènes sont constituées de paralogues et d'orthologues descendant d'un même ancêtre commun. Si au moins une des espèces du clade a plusieurs copies d'un même gène, tous les orthologues et le(s) paralogue(s) de ce gène constituent une famille de gènes. Selon ce classement, toutes les espèces du clade ont alors le même nombre de familles de gènes. Elles possèdent un nombre similaire de gènes impliqués dans ce métabolisme et de gènes en famille. Ces nombres sont résumés dans le tableau 2.4.

Tableau 2.4 : Répartition des gènes du métabolisme lipidique en singletons ou en famille de gènes dans les génomes du clade *Yarrowia* et OLHI : nombre de gènes dans le génome, du nombre de gènes en familles de gènes et du nombre de familles impliquées dans le métabolisme lipidique.

	YALI	**YAGA**	**YAYA**	**YAPH**	**YAAL**	**OLHI**
# Gènes	6642	6936	6282	5916	6209	5338
# Gènes lipidiques	204 (3%)	201 (2,9%)	188 (3%)	176 (3%)	199 (3,2%)	160 (3%)
# Familles lipidiques	37	37	37	37	37	37
# Gènes en famille	136	133	121	109	131	93
# Singletons	68	68	67	67	68	67

YALI possède le plus grand nombre de gènes lipidiques bien qu'elle ne possède pas le plus grand génome. Seuls les gènes CNT2, CNT4, LSC2, PDC1 et SLC1 sont dupliqués dans un autre génome alors qu'ils sont en copie unique chez YALI. CNT2 et CNT4 sont impliqués dans la synthèse de la carnitine ; LSC2 est la sous unité beta du complexe permettant le passage du succinate au succinyl-CoA ; PDC1 permet la formation de pyruvate à partir d'oxaloacétate et SLC1 est impliqué dans la synthèse de DAG. Les gènes présents en une seule copie chez YALI sont également tous présents en au moins une copie dans les autres génomes. Pour chaque espèce, au moins une copie est présente par famille multigénique. Si l'on considère qu'il y a

complémentation fonctionnelle par les gènes d'une même famille au sein d'une même espèce alors toutes les fonctions relatives au métabolisme lipidique existant chez YALI sont également assurées dans les autres espèces du clade et même OLHI.

4 De nouveaux gènes pour le métabolisme lipidique

Les gènes de YALI trouvés dans la bibliographie ou par homologie à SACE sont fournis dans le tableau en Annexe 2. Les annotations présentées dans ce tableaux sont celles disponibles dans Génolevures lorsqu'elles sont correctes, dans le cas contraire je fournis dans ce tableau une proposition pour les nouvelles annotations. Pour les parties du métabolisme lipidique pour lesquelles aucune donnée n'était disponible, je propose ci-dessous des gènes candidats qui sont également présentés dans le tableau en Annexe 2.

4.1 L'alcool gras déshydrogénase FAO

Les alcools gras déshydrogénases FAO catalysent l'oxydation des alcools gras en aldéhydes gras dans les peroxysomes. Cette réaction est utilisée au cours du passage des alcanes aux acides gras (voir Chapitre 2, I, 1.2.1). Chez plusieurs levures comme par exemple *Candida tropicalis*, *Candida cloacae* et *Debaryomyces hansenii* plusieurs FAO ont été isolées et caractérisées (Eirich et al., 2004; Vanhanen et al., 2000) mais aucune FAO n'a été mise en évidence chez YALI. Pour trouver une potentielle FAO dans le clade *Yarrowia* nous avons effectué des blasts de séquences protéiques répertoriées au NCBI dispersées dans tout l'arbre phylogénétique des eucaryotes. Malgré cette recherche très étendue, aucune FAO potentielle n'a été trouvée chez YALI, ni dans les autres espèces du clade, ni chez OLHI. Ceci peut s'expliquer de plusieurs façons. Soit il existe chez ces organismes une nouvelle enzyme capable d'effectuer cette oxydation mais nous ne disposons pas d'homologue nous permettant de la trouver. Soit cette réaction est catalysée par une enzyme proche possédant des domaines similaires tel que les oxydases ERG codées par YALI0E17347g et YALI0F11297g, homologues à ScERG25, et YALI0D20878g, homologue à ERG3. Mais aucun de ces trois gènes ne possède de signal d'adressage au peroxysome. Une autre possibilité serait que le passage de l'alcool gras en aldéhyde gras soit effectué exclusivement dans le réticulum endoplasmique par l'action des alcools déshydrogénases ADH. Actuellement nous n'avons pas suffisamment d'évidences expérimentales pour proposer un gène candidat.

4.2 Les acyl-CoA synthétase FAA

Chez SACE il existe des acyl-CoA synthétases FAA1, FAA2, FAA3 et FAA4 qui interagissent directement avec le transporteur FAT1 (Obermeyer et al., 2007). ScFAT1 permet l'entrée des acides gras dans la cellule ou les peroxysomes et les FAA permettent l'activation de ces acides gras dans la cellule avec ScFAA1, ScFAA4, ScFAA3 et dans les peroxysomes avec ScFAA2 (voir Chapitre 2, I, 1.3). Les transporteurs FAT1 sont membranaires et les FAA sont localisés dans le compartiment concerné, cytosol ou peroxysome, et ne sont pas membranaires. On peut effectivement observer la présence d'un PTS1 sur les séquences codant ScFAA2 et ScFAT1 mais pas pour ScFAA1, ScFAA3 et ScFAA4.

Chez les espèces du clade et OLHI il existe un seul homologue aux FAA de SACE : YlFAA1 YALI0D17864g. YlFAA1 est annoté comme similaire à FAA1 de SACE. Ces deux gènes partagent en effet 68.1% de similarité. Par contre, il n'existe pas d'autre FAA chez les espèces du clade et OLHI. Les autres hits qui sortent lorsque l'on réalise un blast sont les coumarate-CoA ligases COU1 à COU10. Les séquences de FAA1 du clade et de OLHI n'ont pas de motif d'adressage aux peroxysomes, ni PTS1 ni PTS2. Concernant le transporteur FAT1, il existe quatre homologues chez chaque espèce du clade et chez OLHI. Chez YALI les gènes YlFAT1 YALI0E16016g, YlFAT2 YALI0E12859g, YlFAT3 YALI0B05456g et YlFAT4 YALI0C09284g partagent respectivement 60,6%, 30,1%, 32,4% et 32,3% de similarité avec ScFAT1. YlFAT1 est annoté comme similaire à ScFAT1 et YlFAT2, YlFAT3 et YlFAT4 sont annotées comme similaires à ScFAT2, une oxalyl-CoA synthétase capable de convertir l'oxalate en oxalyl-CoA, requise dans la voie de dégradation de l'oxalate. Les séquences des FAT1 du clade et de OLHI possèdent, comme ScFAT1, un motif d'adressage aux peroxysomes de type PTS1. Ceci suggère que ces FAT1 pourraient être requises, comme chez SACE, pour l'entrée des AG dans le cytosol et dans les peroxysomes. De même les FAA1 du clade peuvent réaliser l'activation des AG dans le cytosol mais il n'y a pas de FAA peroxysomal dans le clade. Il existe en revanche dans le clade la famille COU, composée de gènes annotés comme coumarate-CoA ligase chez YALI. Ces enzymes seraient capables de lier un CoA à un acide et toutes les coumarate-CoA ligases du clade ont une séquence d'adressage aux peroxysomes. Nous proposons alors que l'apport des acides gras activés dans les peroxysomes pourrait se faire de deux manières : soit les acides gras sont d'abords activés dans le cytosol puis sont importés dans le peroxysome soit les acides gras entrent dans le peroxysome et y sont activés par les coumarate-CoA ligases.

4.3 L'acétyl-CoA synthétases ACS2 dans les peroxysomes ?

Chez SACE il existe deux acétyl-CoA synthétases, ScACS1 requise en condition aérobie et ScACS2 en condition anaérobie. Elles catalysent la réaction suivante : ATP + acétate + CoA = AMP + diphosphate + acétyl-CoA (voir Chapitre 2, I, 3.1.2). Chez YALI il n'y a qu'un seul gène homologue YlACS2 YALI0F05962g avec 64,2% de similarité avec ACS1 et 62,9% avec ACS2. Cette protéine est déjà connue chez YALI mais aucune expérimentation avec YlACS2 n'a été publiée. Il est prédit qu'ASC2 servirait à l'activation des acides gras, des aldéhydes gras et des DCA dans les peroxysomes mais celle ci n'a pas d'adressage au peroxysome. Comme dans le paragraphe précédant, nous proposons les deux hypothèses suivantes : soit les acides gras sont systématiquement activés auparavant dans le cytosol soit ils peuvent être activés dans les peroxysomes par les coumarate-CoA ligases COU. ASC2 permettrait donc comme chez SACE la formation d'acétyl-CoA à partir d'acétate dans le cytosol.

4.4 L'acyl-CoA thioesterase TES1

Chez SACE il existe une acyl-CoA thioesterase ScTES1 peroxysomale qui permet de réguler les quantités d'acyl-CoA, d'AG libre et de CoASH en hydrolysant l'acyl-CoA en AG libres et CoA (voir Chapitre 2, I, 2.2.9). Cette protéine à été caractérisée chez SACE en 2006 (Maeda et al., 2006) mais n'est pas connue chez YALI. Son homologue, YALI0F14729g, que nous appelons YlTES1 chez YALI existe dans tous les génomes du clade et chez OLHI. ScTES1 et YlTES1 partagent 40,6% de similarité et YlTES1 possède un motif PTS1. Cette protéine était annotée comme similaire à l'acyl-CoA thioesterase peroxysomale d'un champignon pathogène de plantes appelé *Bipolaris maydis* (ou *Cochliobolus heterostrophus*) et le meilleur hit obtenu par blast sur le NCBI est la protéine H072_636 du champignon *Dactylellina haptotylaI,* avec laquelle YlTES1 partage 50,9% de similarité.

4.5 Le transporteur de pyruvate mitochondrial codé par MPC1 et MPC2

La voie permettant la synthèse d'acétyl-CoA à partir du pyruvate dans la mitochondrie n'est pas connue chez YALI. Pour entrer dans la mitochondrie le pyruvate utilise un transporteur identifié en 2012 chez les levures, composé de deux sous-unités codées par les gènes ScMPC1 et ScMPC2

chez SACE (voir Chapitre 2, I, 3.1.1) (Herzig et al., 2012). Dans le génome de YALI on trouve les deux protéines YALI0F00264g avec 64% de similarité avec ScMPC1, annotée comme similaire à une protéine hypothétique de *Neurospora crassa* et YALI0C03223g avec 57,2% de similarité avec ScMPC2, annotée comme similaire à ScMPC3, ohnologue de ScMPC2. D'après Mitoprot, YlMPC1 a une probabilité de 0,88 d'être adressé à la mitochondrie alors que YlMPC2 a une probabilité de seulement 0,32 (Claros and Vincens, 1996). Mais ces protéines forment un complexe donc si l'une d'elles est adressée à la mitochondrie cela pourrait suffire pour que le complexe entier soit adressé dans ce compartiment. Les gènes codant les deux sous-unités MPC1 et MPC2 sont également présents dans tous les génomes du clade et chez OLHI.

4.6 YALI a un complexe BCKDC

Chez les animaux il existe trois complexes de la même famille assurant des fonctions proches dans la mitochondrie : le complexe alpha-cétoglutarate déshydrogénase (KGDC) (voir Chapitre 2, I, 3.3.1), le complexe pyruvate déshydrogénase (PDC) (voir Chapitre 2, I, 3.1.1) et le complexe déshydrogénase de cétoacide-alpha à chaîne ramifiée (BCKDC) (voir Chapitre 2, I, 3.1.3). Le KGDC intervient dans le cycle du TCA pour le passage de l'alpha cétoglutarate en succinyl-CoA, le PDC intervient dans la synthèse de l'acétyl-CoA à partir du pyruvate en décarboxylant le pyruvate et le BCKDC intervient dans la deuxième étape de la dégradation de la leucine pour donner l'isovaleryl-CoA. Ces trois complexes sont formés de trois sous unités E1, E2 et E3. Les sous-unités E1 et E2 de ces complexes catalysent des réactions similaires mais sont spécifiques de chaque complexe alors que E3 sert les trois complexes. Chez SACE les gènes codant E1, E2 et E3 du BCKDC ne sont pas identifiés et n'existent probablement pas, suggérant que le PDC ou le KGDC assure également le rôle du BCKDC. Chez les espèces du clade et OLHI les gènes codant le KGDC sont homologues à ceux de SACE et sont bien adressés à la mitochondrie avec par exemple une probabilité de 0,99 pour YlKGD1, YlKGD2 et 0,95 pour YLLPD1. Par contre les complexes PDC et BCKDC n'ont pas été décrits.

En réalisant un blast des deux gènes codant la sous unité E1 du PDC de SACE, ScPDA1 et ScPDB1 contre les génomes du clade et OLHI, nous trouvons deux hits par espèce pour chacun de ces gènes que nous appelons PDA1, PDA2 et PDB1, PDB2. YlPDA1 YALI0F20702g et YlPDB1 YALI0E27005g sont annotés comme similaires aux sous unités alpha et beta de la pyruvate déshydrogénase de la levure *Candida tenuis.* YlPDA2 YALI0D08690g et YlPDB2 YALI0F05038g sont annotés comme similaires aux sous unité alpha et beta de la 2-oxoisovalerate déshydrogénase du champignon pathogène de plantes *Pyrenophora tritici-repentis*. Elles ont toutes les quatre une probabilité d'être adressée à la mitochondrie de 0,99. La prédiction des GO term de ces quatre protéines montre que YlPDA1 aurait une activité pyruvate déshydrogénase et

oxydoréductase alors que YlPDA2 aurait seulement une activité oxydoréductase. YlPDB1 et YLPDB2 auraient toutes deux une activité catalytique.

Toutes ces observations suggèrent que YlPDA1 et YlPDB1 seraient impliqués spécifiquement dans le PDC et YlPDA2 et YlPDB2 dans le BCKDC pour la dégradation de la leucine et il en serait de même pour toutes les espèces du clade et OLHI.

Par contre les gènes de SACE codant E2 du complexe PDC, ScLAT1, et E3, ScLPD1, possèdent chacun un seul homologue dans le clade et OLHI. Chez YALI ces gènes sont YlLAT1 YALI0D23683g et YlLPD1 YALI0D20768g avec 0,95 de probabilité d'être adressés à la mitochondrie. Mais YALI possède aussi une dihydrolipoamide acétyltransférase YlBKT1 YALI0D23815g homologue à la protéine codant E2 chez l'humain et coderait donc la E2 du BCKDC de YALI. Les autres espèces du clade et OLHI possèdent aussi un homologue à ce gène. La présence de BCKDC est une originalité du clade *Yarrowia* et de OLHI.

Le tableau suivant fournit un résumé des gènes impliqués dans ces trois complexes chez SACE, lorsqu'ils existent, et dans le clade *Yarrowia* (Tableau 2.5).

Tableau 2.5 : Les gènes constituants les sous unités E1,E2 et E3 des complexes alpha-cétoglutarate déshydrogénase (KGDC), pyruvate déshydrogénase (PDC) et cétoacide-alpha déshydrogénase à chaîne ramifiée (BCKDC) dans le clade *Yarrowia* et chez SACE.

	PDC		**KGDC**		**BCKDC**	
	Clade *Yarrowia*	**SACE**	**Clade *Yarrowia***	**SACE**	**Clade *Yarrowia***	**SACE**
E1	PDB1/PDA1	PDB1/PDA1	KGD1	KGD1	PDB2/PDA2	-
E2	LAT1	LAT1	KGD2	KGD2	BKT1	-
E3	LPD1	LPD1	LPD1	LPD1	LPD1	-

4.7 La 3-oxoacid CoA-transférase SKT1

En 2011, Varopreeda *et al.* prédisent que, lors de la fin de la voie de dégradation de la leucine, l'acétoacétyl-CoA provient directement du 3-hydroxy-3méthylglutaryl-CoA par l'action de l'hydroxymethylglutaryl-CoA lyase, la MVA. Mais il manque une réaction intermédiaire. La MVA est capable de produire de l'acétoacétate et de l'acétyl-CoA à partir du 3-hydroxy-3méthylglutaryl-CoA (voir Chapitre 2, I, 3.1.3). L'acétoacétate doit ensuite être activé en acétoacétyl-CoA par une autre enzyme avec une action 3-oxoacid CoA-transferase SKT1. Les protéines qui réalisent cette réaction ont l'EC number : EC 2.8.3.5. Cette protéine a été décrite chez plusieurs bactéries, des eucaryotes et chez l'homme (Bregere et al., 2008). Chez l'homme elle est composée de deux sous unités homologues A et B, codées par deux gènes OXCT1 et OXCT2. En réalisant un blast des deux sous unités de la bactérie *Burkholderia rhizoxinica,* un seul

gène chez YALI est trouvé : YALI0F26587g annoté comme similaire à CA0994|IPF10077 de *Candida albicans.* CA0994|IPF10077 est effectivement annotée par homologie comme une 3-oxoacid CoA-transferase et partagent 74,7% de similarité avec YALI0F26587g. La probabilité d'adressage à la mitochondrie de YALI0F26587g est de 0,98. Ce gène est présent dans tous les génomes du clade et chez OLHI. Nous proposons donc le gène YALI0F26587g comme candidat pour cette étape de la dégradation de la leucine.

4.8 Une seconde TGL3

Deux triacylglycérol lipases intracellulaires, YlTGL3 et YlTGL4, ont été décrites chez YALI (voir Chapitre 2, Introduction, 6.1) (Dulermo et al., 2013). En retravaillant les familles Génolevures comme expliqué dans le paragraphe 3.2, nous avons vu que TGL3 est inclus dans la famille GL3C0731, comprenant un membre chez SACE et deux chez YALI, YlTGL3 YALI0D17534g et YALI0D16379g que nous avons appelé YlTGL32. SACE possède deux autres TGL, ScTGL4 et son ohnologue ScTGL5. YlTGL32 partage 26,6% de similarité avec YlTGL3, 28.2% avec ScTGL3, 31,2% avec YlTGL4, 32,7% avec ScTGL4 et 29,1% avec ScTGL5. L'arbre phylogénétique des gènes TGL3, TGL4, TGL5 et TGL32 de SACE et YALI montre que YlTGL32 ne peut pas être affiliée plus à TGL3 que TGL4 et semble avoir divergée avant la différenciation de TGL3 et TGL4 (Figure 2.17).

YlTGL32 a un homologue dans toutes les espèces du clade et chez OLHI. YlTGL32 est exprimée plus particulièrement en milieu acide oléique mais aussi sur glucose et tributyrine alors que YlTGL3 est exprimée principalement sur acide oléique. YgTGL3 est légèrement exprimée sur glucose et très peu sur acide oléique et tributyrine et YgTGL32 a le même profil mais son expression est plus élevée que celle de YgTGL3. YpTGL3 est exprimée plus particulièrement en milieu acide oléique mais aussi sur glucose et tributyrine et YpTGL32 est exprimée sur les trois milieux.

Nous proposons alors que YlTGL32 joue effectivement un rôle en tant que triacylglycérol lipase. Dans les trois espèces, elle semble peu régulée par le milieu, et est plus exprimée que TGL3, dans nos trois conditions. Il serait donc intéressant d'effectuer des études expérimentales sur ce gène.

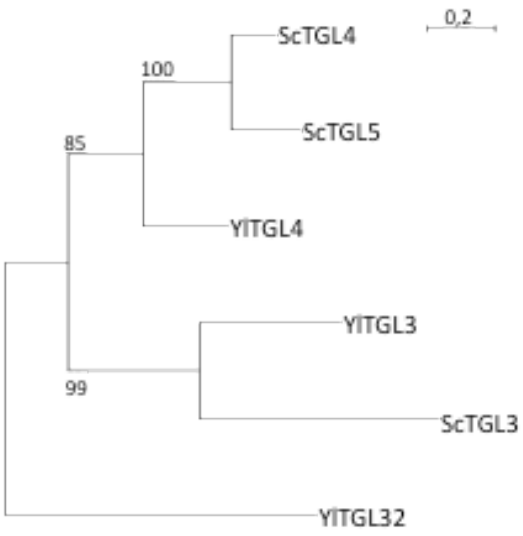


Figure 2.17 : Arbre phylogénétique des gènes TGL3, TGL32, TGL4 et TGL5 de SACE et YALI, 238 résidus.

4.9 La Beta Oxydation mitochondriale

En 2011, Haddouche suggère l'existence d'une ß-oxydation mitochondriale chez YALI comme chez les animaux (voir Chapitre 2, I, 2.2.6) (Haddouche et al., 2011). Il propose comme gènes candidats l'acyl-CoA déshydrogénase YALI0D15708g YlACD2, la 3-hydroxyacyl-CoA déshydrogenase YALI0C08811g et la 3-cétoacyl-CoA thiolase YlPAT1 YALI0E11099g. Il ne propose pas d'homologue mitochondriale à YlMFE1. Cette hypothèse est reprise en 2012 par Varopreeda, qui propose les fonctions candidates EC 1.3.99.-, EC 1.1.1.35, EC 4.2.1.17 et EC 2.3.1.9. KEGG donne les gènes correspondant à tous ces EC number chez YALI sauf pour EC 1.1.1.35 qui est la fonction 3-hydroxyacyl-CoA déshydrogénase. Les gènes sont YlACD2, l'enoyl-CoA hydratase YlECH2 YALI0B10406g et YlPAT1.

Nous avons étudié ces hypothèses et nous proposons également YlACD2 pour la première étape, YlECH2 pour la seconde, YALI0C08811g, que nous décidons d'appeler HBD1, pour la troisième et YlPAT1 pour la dernière. YlECH2 appartient à la famille Génolevures GL3C0501 composée de trois membres chez YALI : YlECH1, YlECH2 et YlMGH1. YlECH1 n'a pas de séquence d'adressage aux mitochondries prédite par MitoProt. Par contre, le gène YlMGH1 YALI0F22121g a 47,6% de similarité avec YlECH2 et a une probabilité de 0,99 d'être adressé aux mitochondries selon MitoProt. Il est annoté comme similaire à l'enoyl-CoA hydratase de *Neurospora crassa* (uniprot|Q7S1U1). Ce gène est proposé par KEGG pour le passage du 3-methyl-glutaconyl-CoA mais nous ne pouvons pas l'exclure comme deuxième candidat pour la seconde étape de la ß-oxydation mitochondriale.

4.10 La néo-synthèse mitochondriale des AG

SACE possède une voie de synthèse des acides gras dans le cytosol par le système de la FASI et une voie dans les mitochondries grâce au système FASII (voir Chapitre 2, I, 5.2) (Brody et al., 1997; Kursu et al., 2013). Le système FASII mitochondrial n'a jamais été décrit chez YALI. Il est composé, chez SACE, d'une protéine malonyl-CoA:ACP transferase codée par ScMCT1 YOR221C, d'une β-céto-acyl-ACP synthase ScCEM1 YER061C, d'une 3-oxoacyl-acyl-carrier-protein réductase ScOAR1 YKL055C, d'une 3-hydroxyacyl-thioester déshydratase ScHTD2 YHR067W et d'une enoyl-ACP réductase ScETR1 YBR026C.

Dans tout le clade et même chez OHLI nous retrouvons les homologues de ces gènes. Chez YALI les homologues sont YlMCT1 YALI0E18590g, YlCEM1 YALI0F30679g, YlHTD2 YALI0A19096g, YlETR1 YALI0C19624g. Par contre pour OAR1 la situation est plus compliquée car ScOAR1 YKL055C appartient à la famille Génolevures GL3C0008 qui compte 10 membres chez SACE et 18 membres chez YALI. Le gène qui a la plus grande similarité avec ScOAR1 est YALI0C19965g. Ce gène est annoté comme ayant quelques similarités avec ScFOX2 alors que l'orthologue de ScFOX2 est YlMFE1 et sa longueur est de 900 aa contre 278 aa pour ScOAR1. Les trois autres gènes similaire à ScOAR1 sont YALI0F29975g et YALI0D05929g annotés comme faiblement similaire à la 3-oxoacyl-acyl-carrier protéine réductase de la bactérie *Chlamydia trachomatis* (uniprot|P38004) et YALI0E05643g annoté comme similaire à l'oxido-réductase de la bactérie *Thermotoga maritima*. Nous avons alors construit l'arbre phylogénétique des hits obtenus avec le blastp de ScOAR1 contre les génomes du clade et OLHI qui ont une e-value inférieure ou égale à $1e^{-8}$ (Figure 2.18). Sur cet arbre, de manière surprenante, une protéine de YAPH, YAPH0S8-03048g, est regroupée avec ScOAR1 mais sa branche est très longue. Les valeurs des bootstraps de cet arbre étant très faible nous pouvons penser que l'emplacement de YAPH0S8-03048g est mal résolu. ScOAR1 et YAPH0S8-03048g partagent 35,5% de similarité mais YAPH0S8-03048g à une probabilité de seulement 0,4 d'être adressé aux mitochondries et ne peut donc pas assurer la même fonction que ScOAR1 chez YAPH. Nous pouvons également observer sur l'arbre l'apparition de deux familles de gènes après la divergence de ScOAR1. La première famille est composée des orthologues de YALI0C19965g annotés contre ScFOX2. Ces gènes possèdent tous une séquence d'adressage aux mitochondries, avec une probabilité selon MitoProt de 0,9 pour tous les gènes, sauf l'orthologue de OLHI, OLHI0E22980g, qui a une probabilité de seulement 0,3 d'être adressé à ce compartiment. La seconde famille comprend les gènes orthologues à YALI0F29975g et YALI0D05929g annotés contre ScOAR1. Selon MitoProt, aucun de ces gènes n'a de séquence d'adressage à la mitochondrie sauf peut être YALI0F29975 avec une probabilité de 0,7. Un autre groupe constitué des gènes orthologues à YALI0E05643g a divergé avant ScOAR1, ces gènes correspondent à SPS19.

Les gènes appartenant à la première famille décrite sont donc des gènes candidats codant potentiellement une 3-oxoacyl-acyl-carrier-protein réductase telle que ScOAR1, avec YALI0C19965g et ScOAR1 qui partagent 36,9% de similarité.

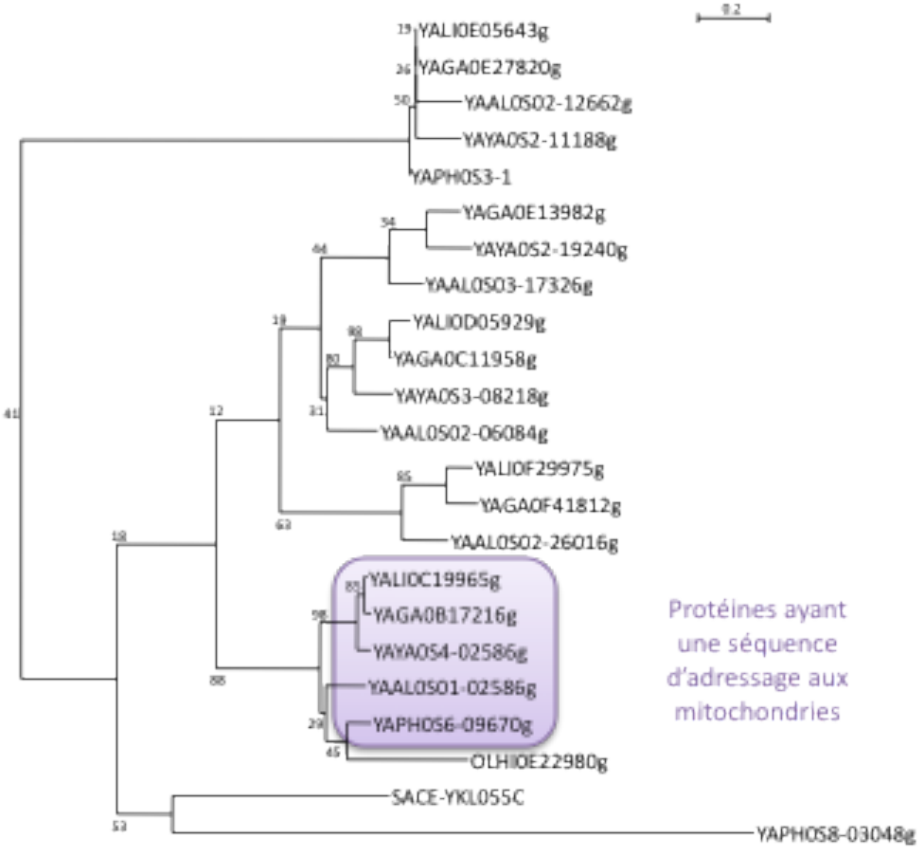


Figure 2.18 : Arbre phylogénétique des hits obtenu avec le blastp de ScOAR1 contre les génomes du clade et OLHI qui ont une e-value supérieure ou égale à $1e^{-8}$. 23 gènes, 112 résidus.

4.11 Les enzymes auxiliaires

Les enzymes auxiliaires assurent les étapes préliminaires permettant aux AG cis- et trans-insaturés ayant des doubles liaisons en position impaire et les AG cis- insaturés en position paire d'entrer dans le cycle de la ß-oxydation (voir Chapitre 2, I, 2.2.6). Chez SACE ces enzymes sont une 2,4-dienoyl-CoA reductase SPS19 (YNL202W), une Δ^3,Δ^2-enoyl-CoA isomérase ECI1 (YLR284C) et une $\Delta^{3,5},\Delta^{2,4}$-dienoyl-CoA isomérase DCI1 (YOR180C).

Chez YALI, l'enzyme SPS19 est connue mais les homologues de ECI1 et DCI1 n'ont jamais été décrits (Mizugaki et al., 1985). Contrairement à ce qui avait été proposé par Mizugaki *et al.*, nous ne trouvons pas une mais 3 copies du gène SPS19 que nous appelons SPS191 pour YALI0D01694g, SPS192 pour YALI0E14322g et SPS193 pour YALI0C03003g et qui partagent de 76% à 88% de similarité. SPS193 est présent dans tous les génomes du clade et chez OLHI. Par contre SPS192 n'est pas retrouvé chez OLHI et SPS191 n'est présent que chez YALI, YAGA et YAAL. Les SPS19 sont des protéines très conservées dans le clade, avec des pourcentages de similarité allant de 79% pour YgSPS193-OhSPS193 à 98% pour YlSPS192-YgSPS192. Elles

possèdent toutes une séquence N-terminale très proche du motif consensus d'adressage aux peroxysomes PTS1 : PKI, SKI ou PKL, suggérant qu'elles sont probablement bien adressées aux peroxysomes, excepté pour YgSPS193 YAGA0F11958g. Mais lorsque l'on regarde la séquence de YgSPS193 on retrouve la séquence N-terminale homologue aux autres SPS19 du clade sur une autre phase de lecture que celle de la CDS car la séquence posséde 6 bases C au lieu de 5 (Figure 2.19). Il n'y a pas de problème dû au séquençage car le codon stop est confirmé par le séquençage des ARNm. Ce codon stop et le frameshift résulte peut être d'une mutation ponctuelle sur cette souche. Cette mutation induit une région C terminale différente ou nécessite un frame shift programmé pour lire la phase correspondante chez les autres SPS19 du clade.

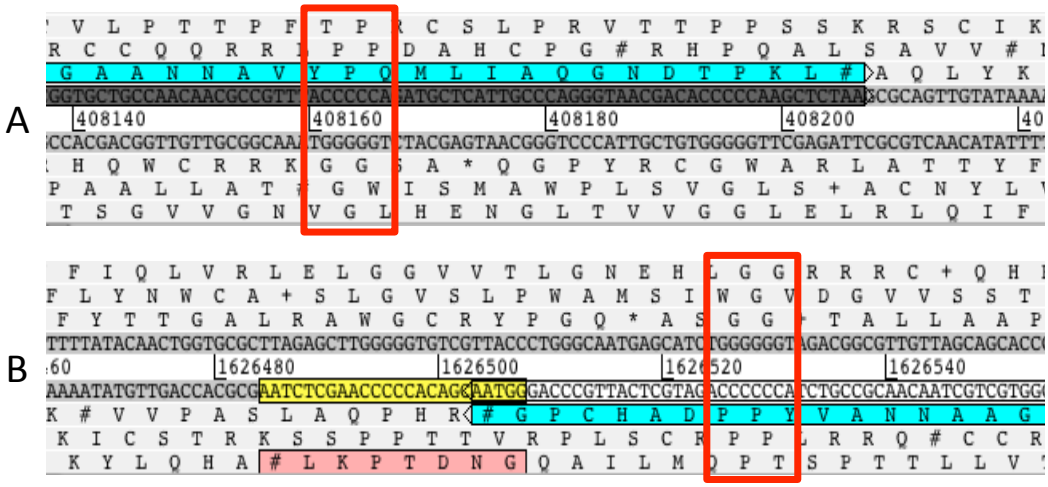


Figure 2.19 : Régions N-terminale des gènes SPS193. A) YlSPS193 présentant un potentiel motif d'adressage PTS1 avec la séquence terminale PKL. B) YgSPS1913 sans motif PTS1, avec une séquence identique à la région N-terminale de YlSPS193 sur une autre phase de lecture, suggérant une mutation antérieure ayant abouti a un frameshift. La mutation est une base supplémentaire au niveau de l'encadré rouge.

Concernant les gènes ScDCI1 et ScECI, il n'y a pas d'homologue chez YALI, ni dans le clade, ni dans OHLI. En réalisant un blastp de ces gènes contre les génomes du clade et OHLI, les hits trouvés sont trop éloignés et trop courts. Nous avons donc recherché une $\Delta^{3,5},\Delta^{2,4}$-dienoyl-CoA isomérase et une Δ^{3},Δ^{2}-enoyl-CoA isomérase dans ces génomes qui ne seraient pas homologues à celles de SACE.

Il existe chez YALI un gène, YALI0A07733g, annoté comme similaire à *Neurospora crassa* uniprot|Q8WZH4, qui serait une possible $\Delta^{3,5},\Delta^{2,4}$-dienoyl-CoA isomérase. Ce gène YlECH1 a deux paralogues YlECH2 YALI0B10406g et YlMGH1 YALI0F22121g. YlECH2 et YlMGH1, qui ont des séquences d'adressage aux mitochondries, sont les gènes candidats proposés dans cette étude pour la seconde étape de la ß-oxydation mitochondriale (voir paragraphe 5.10). ECH1 est présent dans tous les génomes du clade et chez OLHI et n'a pas d'homologue chez SACE. Ces ECH1 présente une séquence N-terminale, EKL, très proche du motif consensus des PTS1 et serait donc bien adressée aux peroxysomes.

Pour trouver un gène codant une Δ^3,Δ^2-enoyl-CoA isomérase dans le génome de YALI nous avons recherché des enoyl-CoA isomérases dans des organismes plus éloignés. Chez les champignons tels que *Aspergillus oryzae*, *Aspergillus niger, Penicillium oxalicum, Penicillium digitatum* et autres *Aspergillus* ou *Penicillium*, ces protéines appartiennent à des familles multigéniques. Chez YALI il existe une famille dont les membres sont homologues aux énoyl-CoA isomérases de ces champignons filamenteux. Cette famille est composée de 5 membres et correspond à la famille Génolevures GL3R3833 qui compte des membres seulement chez YALI. Sachant que YALI est l'espèce la plus basale dans l'arbre des espèces disponibles sur Génolevures ceci pourrait suggérer que ces gènes ont été conservés depuis la divergence des champignons puis perdus après la divergence de YALI ou bien qu'il y a eu un transfert horizontal d'un des gènes de la famille d'un champignon vers un ancêtre de YALI. Nous avons décidé d'appeler cette famille ISO, avec les membres ISO1 YALI0F28567g, ISO2 YALI0D09493g, ISO3 YALI0A20207g, ISO4 YALI0D09383g et ISO5 YALI0D00671g. Les orthologues sont retrouvés dans tout le clade, sauf ISO1 et ISO4 absents chez YAPH. OLHI possède seulement les gènes IS02 et ISO5. Toutes les ISO1, ISO2 et ISO3 possèdent un PTS1 AKL. Les ISO5 possèdent un PTS1 SKL ou TKL pour OLHI ou GKL pour YAPH. Les ISO4 par contre ne possèdent pas de séquence caractéristique des PTS1 et PTS2, elles ne sont probablement pas adressées aux peroxysomes. Nous proposons donc comme candidats codants les enzymes auxiliaires les gènes SPS191, SPS192, SPS193, ECH1, ISO2 et ISO3.

4.12 Cas d'évolution de l'épissage alternatif

Parmi les gènes impliqués dans le métabolisme lipidique, certains présentent soit plusieurs départs alternatifs de transcription soit un épissage alternatif de leurs introns. La malate déshydrogénase MDH2 de YALI est alternativement épissée grâce à la variation de la taille du second intron (Kabran et al., 2012). Cet épissage alternatif produit une forme de malate déshydrogénase peroxysomale avec une séquence C-terminale contenant un motif d'adressage PTS1 et une autre forme cytosolique sans motif d'adressage. Les deux formes, peroxysomale et cytosolique, se retrouvent dans tous les génomes du clade et chez OLHI.

Un second cas d'épissage alternatif est trouvé chez YALI, YAGA, YAYA et YAAL. Une des deux formes du gène HFD2, codant une déshydrogénase d'aldéhydes gras (voir Chapitre 2, I, 1.2.1), est constituée d'un seul exon tandis que la seconde forme a un intron dans la région C-terminale et un second exon traduit sur une autre phase de lecture (Figure 2.20). La séquence de la forme à intron a un motif d'adressage aux peroxysomes de type PTS1 alors que la forme longue n'a pas de motif d'adressage et semble donc être cytosolique. Chez YAYA, le gène HFD2 est dupliqué et une des deux copies fait de l'épissage alternatif comme décrit précédemment tandis que la seconde copie, YAYA0S5-00760g, n'a pas d'intron et semble être peroxysomale avec un motif

PTS1. Le gène HFD2 appartient à une famille de 4 membres chez YALI, YAGA et YAYA, à 3 membres chez YAAL et 2 chez YAPH et OLHI. YAPH et OLHI ne possèdent pas le gène HFD2. Les autres gènes de la famille, HFD1 et HFD3, ne possèdent pas d'adressage aux peroxysomes chez YAPH. Chez OLHI, par contre, le gène HFD1 peut, comme HFD2, être épissé alternativement et la forme à intron a un adressage aux peroxysomes.

Le dernier cas d'épissage alternatif détecté au cours de cette étude concerne le gène codant l'acétyl-CoA carboxylase ACC1 chez YAAL, YAAL0S05-01332g (voir Chapitre 2, I, 5.2) (Figure 2.20). Ce gène est composé de 3 exons. Le premier intron possède deux sites d'épissage en 3' consécutifs. L'épissage de cet intron génère deux isoformes ne différant que par 3 nucléotides. Nous pouvons donc nous demander si cet épissage alternatif a une incidence sur la fonction ou l'activité de la protéine.

Il existe également plusieurs cas de départs alternatifs dans les espèces du clade *Yarrowia*. Chez YALI deux formes du répresseur de transcription des gènes ALK codé par YAS3 sont produites par un départ alternatif de transcription (Figure 2.20) (Hirakawa et al., 2009). L'une des formes est composée de deux exons et d'un intron. La transcription de la seconde forme, composée d'un seul exon, démarre dans le second exon de la première forme. La même organisation est retrouvée dans toutes les espèces du clade sauf chez YAAL et n'est pas retrouvée chez OLHI. Chez YAAL et OLHI seule la forme longue à deux exons existe. Ceci suggère que la forme courte de YAS3 est apparue après la divergence de YAAL.

Le gène EDH3, codant une énoyl-CoA hydratase, possède deux formes chez YAPH (Figure 2.20). Une forme longue est composée de deux exons et la forme courte démarre quelques acides aminés après le début du second exon de la forme longue. En amont de la forme courte il y a un intron en 5'UTR dans l'ARN messager. Ce gène possède donc deux départs de transcription et un intron spécifique à chaque forme. Chez YALI et YAGA, il également y a deux départs de transcriptions différents avec chacun un intron spécifique en 5'UTR mais ils génèrent la même CDS qui correspond à la forme courte de YpHFD3. Chez YAYA, YAAL et OLHI, seule la forme courte existe et il n'y a qu'un seul départ de transcription.

Chez YAYA nous retrouvons deux isoformes pour le gène FUM1, codant une fumarase intervenant dans le cycle du TCA (voir Chapitre 2, I, 3.3.1). Il existe une forme longue avec un intron et une forme courte sans intron qui démarre à la même position que le second exon de la forme longue. En amont de la forme courte il y a un intron en 5'UTR donc la forme courte et la forme longue ont le même départ de transcription mais les CDS sont différentes. Chez les autres espèces du clade et OLHI seule la forme courte à intron en 5'UTR existe.

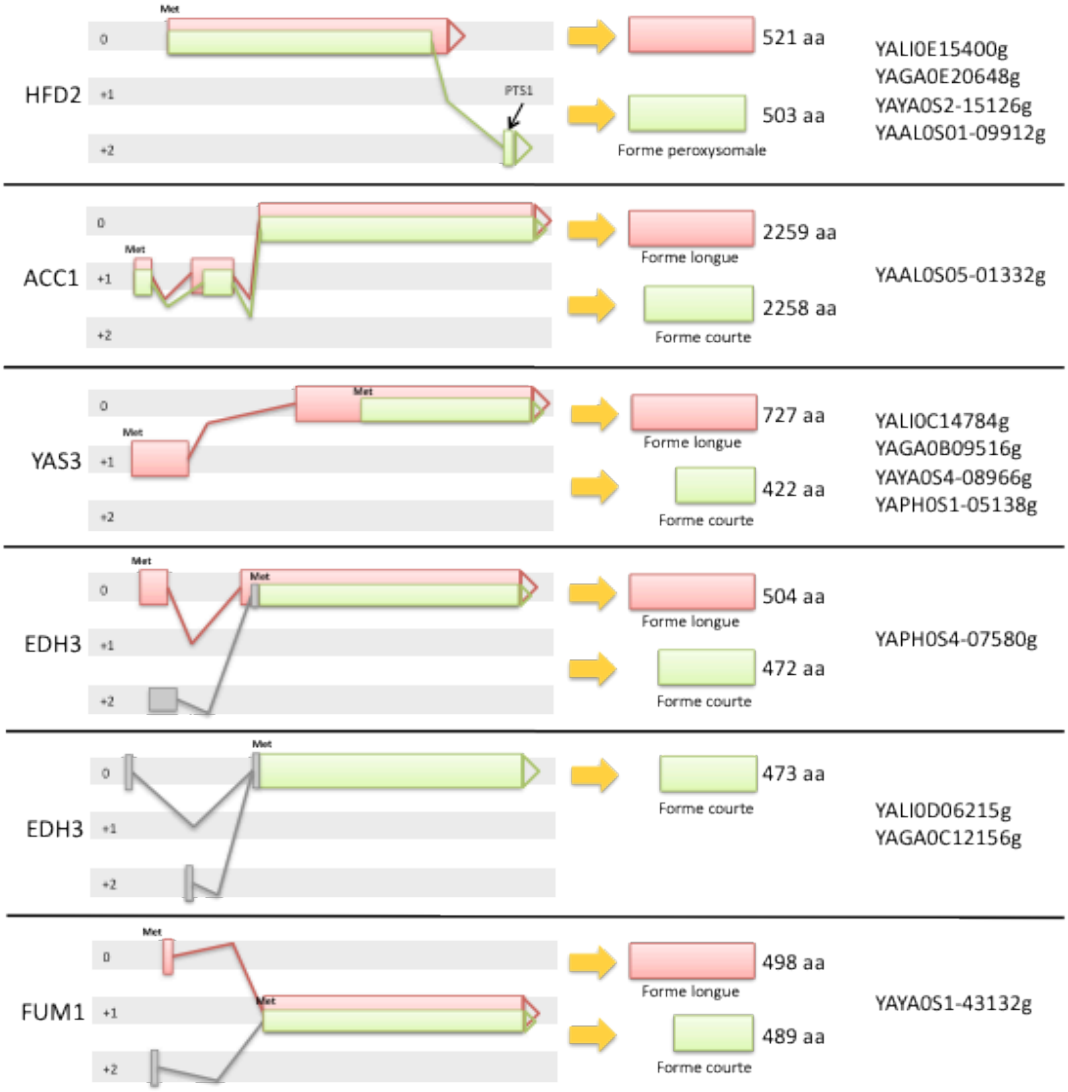


Figure 2.20 : Représentations schématiques des différents transcrits alternatifs provenant des gènes HFD2, ACC1, YAS3, EDH3 et FUM1. Les exons sont représentés par des rectangles de couleur, rouge pour une première forme et vert pour la seconde forme. Les introns sont représentés par des traits rouges ou verts. Les ARN messager sont représentés en gris et positionnés de façon aléatoire sur un des cadres de lecture. Les barres grises représentent les trois cadres de lecture 0, +1, +2. Les tailles des CDS fournies correspondent à celles de YALI par défaut, lorsqu'il n'y a pas de transcrit alternatif chez YALI la taille correspond à l'espèce concernée.

Le dernier cas de départ alternatif de transcription trouvé est celui du gène ABC5 codant un transporteur d'alcanes (voir Chapitre 2, I, 1.2.1). Le gène ABC5 a deux départs de transcription chez YALI, YAGA et YAAL générant ainsi une forme longue et une forme courte (Figure 2.21). En amont de la forme courte il y a un intron en 5'UTR. Chez YALI et YAGA, les données de transcriptomique permettent de valider ces deux formes avec une couverture importante juste en amont de la forme longue. La forme courte semble par contre minoritaire avec peu de couverture en amont de l'intron en 5'UTR. Sachant que les motifs d'adressage à la mitochondrie

se trouvent en N-terminale des protéines, nous avons regardé s'il existait une différence d'adressage entre les deux formes en utilisant Mitoprot. Aucune des deux formes n'a de motif d'adressage à la mitochondrie.

Chez YAYA, YAPH et OLHI, seule la forme courte de ABC5 existe, toujours avec l'intron en 5'UTR sauf chez OLHI. Ceci suggère que le départ alternatif est apparu après la divergence de YAYA. Par contre chez OLHI et YAAL il existe une seconde copie du gène ancestral de ABC5, que nous appelons ABC5b, qui a été perdu avant la divergence de YAPH. Cette seconde copie a environ 30 acides aminés de moins que la forme courte de ABC5 et elle est composée de deux exons avec l'intron situé juste après l'ATG de départ. La perte de ce gène après la divergence de YAAL pourrait potentiellement avoir été comblée par l'apparition du départ alternatif de transcription chez YALI et YAGA. Il faudrait déterminer la localisation et la spécificité des différents produits des gènes ABC5 et ABC5b et de leurs différentes formes pour pouvoir tester cette hypothèse.

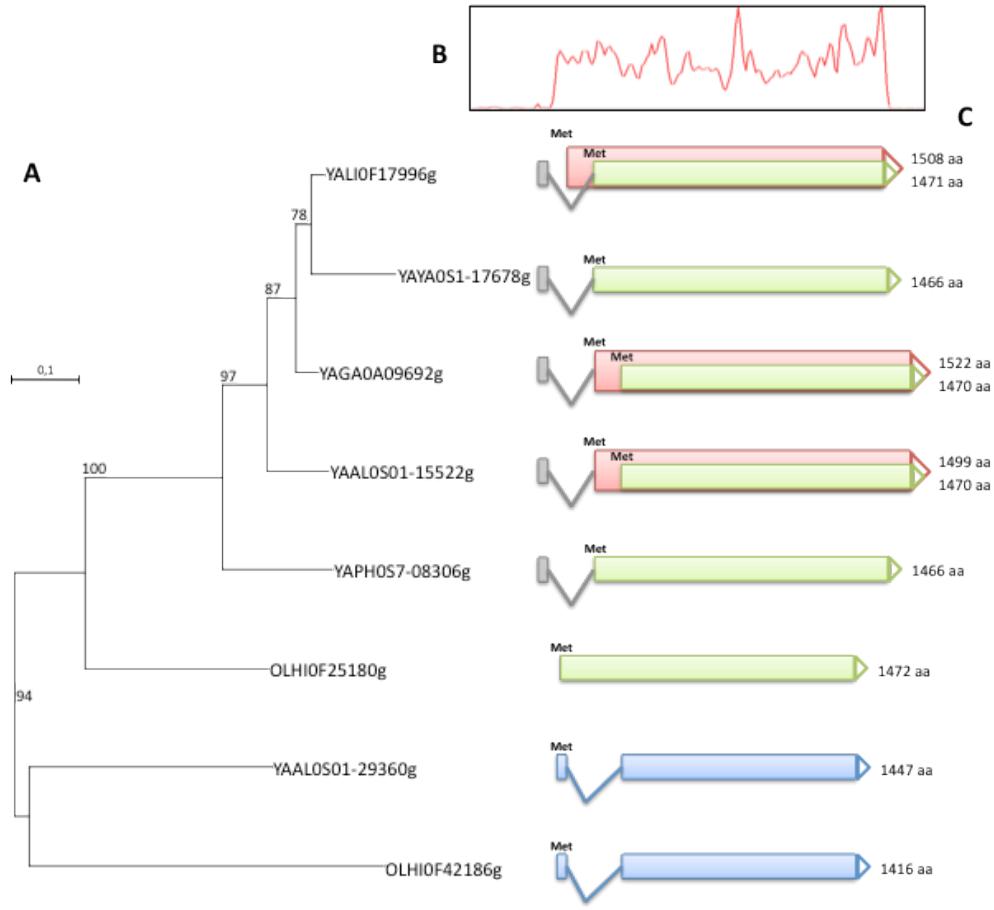


Figure 2.21 : A) Arbre phylogénétiques des gènes ABC5 du clade *Yarrowia* et OLHI, 1490 résidus. B) Couverture des transcrits obtenue en RNAseq sur acide oléique chez YALI. C) Représentations schématiques des différents transcrits alternatifs de ABC5 chez YALI, YAGA et YAAL et des différents transcrits pour les ABC5 et ABC5b des autres espèces. Les exons sont représentés par des carrés de couleur. Les introns sont représentés par des traits. Les ARN messagers sont en gris. Les tailles des CDS sont fournies à côté de la forme correspondante.

IV. CONCLUSION

Pour cette partie de ma thèse, nous voulions principalement savoir si les espèces du clade *Yarrowia* possèdent un métabolisme lipidique similaire, c'est à dire si elles présentent les mêmes voies pour ce métabolisme et si toutes les fonctions sont toujours assurées par des enzymes homologues. Pour répondre à ces questions nous avons dû reconstruire le métabolisme lipidique chez YALI puis chez les autres espèces du clade. Pour cela nous disposions de 5 nouveaux génomes séquencés et annotés dans l'équipe.

Un nouveau genre pour OLHI

Le bilan du séquençage de quatre espèces du clade *Yarrowia* est très intéressant car il nous permet de constater un gradient dans la taille des génomes, de 16,2 Mb pour YAPH à 22,8 Mb pour YAGA, dans le nombre de CDS, de 5916 pour YAPH à 6936 pour YAGA, et également pour le pourcentage en GC, d'environ 49% chez YALI, YAGA, YAYA, YAAL et de seulement 43% chez YAPH. Ces espèces sont relativement proches puisqu'elles partagent de 80% à 91% de similarité entre elles. Par contre le séquençage de OLHI révèle un organisme très divergent autant pour la structure de son génome que pour son contenu en gènes. En effet, son génome de 10,7 Mb est deux fois plus petit, il ne contient que 5338 gènes et son pourcentage en GC est de seulement 41%. La perte totale de synténie entre son génome et celui des autres espèces du clade, son positionnement basal sur l'arbre phylogénétique et son faible pourcentage de similarité avec les autres espèces, en moyenne 67%, sont d'autres arguments aboutissant à exclure OLHI du clade *Yarrowia*. Nous avons donc pu assigner l'espèce *Oleozyma hispaniensis* OLHI, que nous appelions auparavant *Candida hispaniensis*, à un nouveau genre grâce à la caractérisation de son génome. L'étude de son génome sera publié séparément dans un article actuellement en cours de rédaction : « Whole genome sequencing as an help to reclassify *Candida* species: the case of *Oleozyma* (*Candida*) *hispaniensis* » (Michely et al., in prep.).

D'autre part, la différence de taille observée entre les génomes du clade *Yarrowia* et celui de OLHI suggère qu'il y a une expansion des génomes dans le clade qui semble être reliée à la taille moyenne des régions intergéniques. Il serait très intéressant d'étudier plus spécialement les 6 génomes afin de caractériser les mécanismes à l'origine de cette expansion ainsi que les forces agissant pour la conservation de ces grandes séquences intergéniques.

Le métabolisme lipidique

La reconstitution du métabolisme lipidique chez YALI a permis de recenser 204 gènes chez YALI dont 136 sont répartis dans 37 familles de gènes. Cette reconstitution a permis de retrouver tous ces gènes chez les 5 espèces YAGA, YAYA, YAPH YAAL et OLHI. Nous avons constaté que toutes les fonctions nécessaires à ce métabolisme semblent être assurées par au moins un gène dans chaque génome. Tous les gènes présents en une seule copie chez YALI sont également présents en au moins une copie dans les autres espèces et au moins un duplicat est présent par famille multigénique de YALI. Si l'on considère que les gènes d'une même famille assurent la même fonction ou une fonction très proche nous pouvons considérer qu'aucune étape du métabolisme lipidique n'est manquante dans chacune des espèces. Aucune voie métabolique n'est absente dans ces espèces même les voies qui apparaissent comme secondaires telles que la ß-oxydation mitochondriale ou la néo-synthèse mitochondriale des AG. Il a en effet été montré qu'après l'inactivation chez YALI de la YlMFE1, codant l'enzyme multifonctionnelle impliquée dans deux étapes de la ß-oxydation peroxisomale, une activité résiduelle de dégradation est observée dans les souches (Haddouche et al., 2011). La présence de toutes les voies du métabolisme lipidique est plus spécialement surprenante chez OLHI qui a dans son génome environ 1300 gènes de moins que les espèces du clade. La présence de gènes permettant d'assurer toutes les fonctions est en corrélation avec ce que nous avons observé dans le Chapitre 1, en physiologie comparative puisque nous n'avons observé que peu de différences pour les capacités de croissance en fonction des milieux, ainsi que pour le stockage et la synthèse des acides gras.

La génomique comparative pour déterminer de nouveaux gènes

Après avoir reconstruit le métabolisme lipidique grâce aux données provenant de la bibliographie et aux orthologies avec SACE nous avons pu proposer des gènes candidats pour les fonctions manquantes en utilisant la génomique comparative. Ces gènes permettraient de compléter plusieurs voies du métabolisme lipidique : la ß-oxydation mitochondriale, la néo-synthèse mitochondriale des acides gras, l'activation des acides gras dans les peroxysomes, une partie de la régulation des quantités d'acyl-CoA, le transport de pyruvate dans la mitochondrie, la dégradation de la leucine, les étapes préliminaires à la ß-oxydation peroxysomale pour certains AG. Il serait maintenant intéressant de tester tous ces gènes en laboratoire, en les délétant ou les surexprimant chez YALI afin de confirmer leur fonction.

Nous avons aussi montré qu'il existe un complexe BCKDC dans les espèces du clade et OLHI. Dans les tissus animaux le complexe BCKDC catalyse une réaction irréversible dans le catabolisme des acides aminés à chaîne ramifiée L-isoleucine, L-valine, et L-leucine (Harris et al., 1986). Chez l'humain, ce complexe est caractérisé comme l'enzyme de régulation la plus

importante dans le catabolisme de ces acides aminés (Harris et al., 2005). Chez les bactéries il participe à la synthèse des acides gras à longue chaîne ramifiée (Lennzrz, 1961) et chez les plantes, cette enzyme est impliquée dans la synthèse des hydrocarbures ramifiés à chaîne longue (Bharti et al., 2014). Nous ne trouvons aucun article dans la bibliographie décrivant ce complexe chez une levure. De plus, nous pouvons observer grâce à un blastp qu'aucun *Saccharomycetaceae*, ni CTG ne semblent posséder d'homologues des gènes composant le BCKDC.

De façon générale, parmi les 30 nouveaux gènes proposés chez YALI, 9 sont plus proches de gènes de champignons que de leurs homologues chez les hémiascomycètes. Ces gènes sont TES1, MPC1, ECH1, ISO1 à ISO5 et MGH1. Ceci est en corrélation avec la position basale du clade *Yarrowia* dans la phylogénie des Ascomycètes. Ces gènes ont probablement été gardés dans le clade puis perdus après la divergence de YALI mais nous ne pouvons pas exclure l'hypothèse de transferts horizontaux de ces gènes de champignons à l'ancêtre du clade.

Au cours de la reconstruction du métabolisme lipidique nous avons observé la présence de nombreuses familles de gènes impliqués dans ce métabolisme. Les gènes en familles représentent en effet 67% des gènes lipidiques chez YALI. Nous avons pu constater que cette répartition en familles est similaire pour les autres génomes du clade et même chez OLHI bien que ses familles soit moins grandes. De plus, il était déjà connu que les compétences oléagineuses sont facilitées chez YALI par la présence de grandes familles de gènes impliquées dans le métabolisme lipidique avec par exemple la famille des 16 lipases, celle des 10 cytochromes P450 et les 6 acyl-CoA oxydases (Barth et al., 2013). De même nous savons que chez SACE, ces familles ne possèdent qu'un membre. Dans ce contexte, il nous a paru très intéressant d'étudier plus finement ces familles de gènes et leur évolution dans le clade afin de comprendre comment s'est fait cette expansion des familles.

CHAPITRE 3 :

L'ÉVOLUTION DES GÉNOMES ET DU MÉTABOLISME

I. LES OBJECTIFS

Nous avons vu précédemment que les génomes du clade *Yarrowia* possédaient entre 176 et 204 gènes impliqués dans le métabolisme lipidique ce qui représente environ 3% des gènes. Ce métabolisme est primordial pour l'adaptation des levures oléagineuses à leur milieu. Ceci suggère que les séquences des protéines composant ce métabolisme sont soumises à des forces évolutives qui conduiraient soit à la sélection de la divergence dans un processus d'adaptation soit à la contre sélection de nouveautés si le métabolisme est déjà efficace dans un milieu donné.
Le premier objectif de cette partie sera donc de caractériser la conservation et la contrainte sélective agissant sur chacun des gènes lipidiques afin de repérer s'il existe des voies du métabolisme lipidique qui évolueraient différemment. Le second objectif sera de comparer les pressions de sélection agissant sur ce métabolisme avec celles agissant sur le reste du génome. Ces différences pourraient alors permettre d'observer des mécanismes d'adaptation pour ce métabolisme. Enfin, l'évolution des familles des gènes sera plus particulièrement étudiée car elles représentent plus de 65% des gènes de ce métabolisme. Les principales questions posées sont donc :

- Les pressions de sélection agissant sur le métabolisme lipidique sont elles les même que celles qui s'exercent sur les génomes complets ?
- Existe-t-il des voies du métabolisme lipidique qui auraient des pressions de sélection de type ou de force différente ?
- Comment agit la sélection dans les familles de gènes nombreux et dans les familles dynamiques ?

II. INTRODUCTION

Un métabolisme tel que celui des lipides est primordial pour l'adaptation d'une levure oléagineuse à son milieu. On peut donc supposer que les gènes constituants ce métabolisme sont soumis à des forces de sélection de grande amplitude pouvant agir dans les deux sens. S'il y a sélection positive sur des mutations avantageuse alors les gènes peuvent se perfectionner ou évoluer vers de nouvelles fonctions. Si par contre il y a contre sélection des mutations ceci suggèrerait qu'il est important pour le génome de conserver ce gène. Ces évolutions adaptatives sont facilitées lorsque les gènes appartiennent à une famille multigénique car une des copies peut évoluer indépendamment vers une nouvelle fonction. Il est connu que YALI et YAGA ont subi une extension des familles de gènes impliquées dans le métabolisme des lipides avec 18 2-4 dienoyl CoA réductases, 16 lipases, 12 cytochrome P450 CYP52, 11 sphingomyelinases, 6 acyl-CoA oxydases et 14 transporteurs d'acides gras (Barth et al., 2013). Ces familles sont très dynamiques dans le clade, avec de récentes pertes et gains de gènes. La duplication de gènes, à la base de la création de familles multigéniques, et la sélection sont alors de puissants mécanismes évolutifs car elles entrainent une diversification et/ou perte différentielle de gènes.

Dans ce chapitre nous nous intéressons alors aux mécanismes évolutifs, que sont les mutations géniques et plus particulièrement la duplication de gènes et la sélection naturelle, permettant une adaptation évolutive des gènes et des familles de gènes.

1 EVOLUTION DES GÈNES ET SÉLECTION

Les gènes présents au sein des génomes peuvent évoluer, c'est à dire se modifier du fait des forces évolutives, par des mécanismes aléatoires (mutations, recombinaisons …) et d'autres non aléatoires (sélection naturelle, sélection artificielle…). La modification de ces gènes va entraîner l'apparition de plusieurs séquences nucléotidiques très similaires mais avec de petites variations au sein d'une population, les allèles. Si un allèle est bénéfique, soit pour la reproduction soit parce qu'il augmente les chances de survie, sa fréquence va augmenter dans la population. On parle alors d'évolution adaptative ou sélection positive. La sélection adaptative agit au niveau moléculaire du gène. L'identification des gènes ou des régions géniques sujettes à

la sélection positive peut permettre de prédire les gènes qui ont une grande importance au niveau fonctionnel. Il existe deux manières d'évaluer l'évolution adaptative : la première s'intéresse à la divergence des gènes entre orthologues de différentes espèces ou entre paralogues d'un génome et la seconde évalue la fréquence de polymorphisme entre différentes souches d'une même espèce. La divergence peut être évaluée par l'alignement des séquences permettant de déduire leurs degrés de similitude. Mais pour connaître l'amplitude et la direction des pressions de sélection agissant sur les gènes du métabolisme, il faut s'intéresser à la comparaison des mutations nucléiques synonymes et non synonymes.

Le but de ce projet de thèse étant d'évaluer l'évolution du métabolisme lipidique au sein du clade *Yarrowia*, nous nous sommes intéressés exclusivement à la divergence des gènes constituant ce métabolisme par comparaison des orthologues des différentes espèces et des paralogues d'une même espèce.

1.1 Les substitutions affectant un gène

1.1.1 Les types de mutations

Durant la réplication les séquences d'ADN sont normalement copiées à l'identique grâce à la grande fidélité des ADN polymérases (qui sont au nombre de 5 chez E. coli, 9 chez S. cerevisiae et plus de 10 chez l'humain (Kawasaki and Sugino, 2001; Miyabe et al., 2011). Cependant il arrive que des erreurs se produisent lors de la réplication ou bien de la réparation de l'ADN, donnant ainsi naissance à de nouvelles séquences. Dans des conditions environnementales non stringentes, la fréquence effective des erreurs durant la réplication correspond à une base incorrecte pour 10^9 à 10^{10} nucléotides incorporés. Ces erreurs sont alors nommées des mutations nucléiques et sont définies comme le changement d'une base de sorte que le mutant et le type sauvage coexistent dans une population. Malgré leur faible taux, les mutations nucléiques sont la meilleure source de variabilité et de nouveauté dans l'évolution.
Ces mutations peuvent être classées par le type de changement occasionné par l'évènement :
- La recombinaison est l'échange d'une séquence avec une autre,
- La délétion est la suppression d'un ou plusieurs nucléotides dans une séquence,
- L'insertion est l'ajout d'un ou plusieurs nucléotides dans une séquence,
- L'inversion est la rotation de 180° d'un segment du double brin d'ADN d'au moins deux paires de bases,
- Et la mutation par substitution est le remplacement d'un nucléotide par un autre.

Dans cette étude nous voulons comparer le taux de mutation synonymes et non synonymes, nous allons donc nous intéresser seulement aux mutations par substitution.

1.1.2 Le pattern de mutation

Les mutations par substitution sont de deux sortes : les transitions et les transversions. Les transitions sont les changements entre les purines ou entre les pyrimidines (A↔G, C↔T) alors que les transversions sont les changements entre une purine et une pyrimidine (T↔A, T↔G, C↔A, C↔G) (Gojobori et al., 1982). Les transitions sont généralement plus fréquentes que les transversions malgré le fait que l'on ait deux fois plus de transversions possibles (Wakeley, 1996). Elles représentent 60% à 70% de toutes les mutations. Ceci est traditionnellement attribué au biais de transition selon lequel les transitions ont plus de chances de se produire et moins de chances d'être réparées que les transversions. Il faut cependant noter que ce biais n'est pas universel, il semble différer en fonction des espèces et peut même être absent, ce qui est le cas chez *Podisma pedestris* pour lequel il n'y a pas de différence significative entre les taux de transitions et de transversion (Keller et al., 2007).

Ce biais est observable sur tout le génome : régions codantes et non codantes. Il s'explique de manière intuitive sur les régions codantes. En effet, sur les régions codantes, nous comprenons bien que même si les transitions comme les transversions peuvent changer la composition en acides aminés de la protéine, les différences biochimiques dans la protéine tendent à être bien plus importante dans le cas des transversions (Zhang, 2000). Par conséquence la sélection purifiante contre les transversions sera beaucoup plus forte. Mais la sélection va également favoriser les systèmes de réparation de l'ADN qui permettent d'éliminer au maximum les transversions ce qui va s'appliquer au génome complet. Ces deux mécanismes de sélection vont alors affecter le profil de substitution observé dans les régions codantes comme dans les régions non codantes (Keller et al., 2007). De plus, en 2010, il a été également montré chez les levures, que la sélection naturelle agit pour conserver l'organisation fonctionnelle de la chromatine ce qui contribue aussi significativement à l'apparition d'un biais de transition sur l'ensemble du génome (Babbitt and Cotter, 2010).

1.1.3 Les substitutions sur les régions codantes

Les mutations par substitution dans les régions codantes une protéine peuvent être classées selon deux types en fonction de leur impact sur la protéine. Une mutation par substitution est dite synonyme si elle ne change pas l'acide aminé dans la séquence, ce qui est permis par la redondance génétique. Ce type de mutation est aussi appelé mutation silencieuse lorsqu'elle n'est pas perceptible au niveau des acides aminés. Mais toutes les mutations synonymes ne sont pas silencieuses. Certaines peuvent aussi avoir un impact sur la structure de la protéine en désactivant le site de repliement ou, comme il a été montré sur la protéine de ß globine, en activant un nouveau site de repliement (Goldsmith et al., 1983).

L'autre cas de mutation par substitution est la mutation non synonyme qui va alors donner lieu au remplacement d'un acide aminé par un autre. Ces mutations sont dites faux-sens si le nouveau codon ne code pas pour le même acide aminé que le codon précédent. Si par contre le nouveau codon est un codon stop la mutation est dite non-sens. Dans ce cas la traduction est arrêtée prématurément, et la protéine qui en résulte n'est pas fonctionnelle. Sachant qu'il y a 61 codons différents et que chaque codon peut muter en 9 autres codons en changeant un seul nucléotide il y a 61*9 = 549 mutations par substitution possibles. La quasi-totalité des mutations par substitution en première et seconde position sont des mutations faux-sens ou non-sens. Mais grâce à la redondance du code génétique 70% des mutations par substitution sur la troisième base des codons sont des changements synonymes n'affectant pas la séquence protéique et ont beaucoup plus de chances d'être fixées. Ces codons qui codent un même acide aminé sont alors appelés codons synonymes. Si on admet que toutes ces mutations synonymes sont sous sélection neutre et que le patron de mutation est neutre, les codons synonymes devraient être utilisés plus ou moins à la même fréquence. Cependant, les nombreuses études réalisées montrent clairement que l'usage des codons synonymes ne se fait pas de manière aléatoire (Sharp et al., 2010). Chez beaucoup de levures ce biais est remarquable, voir par exemple (Hiraoka et al., 2009) pour *Schizosaccharomyces pombe* ou (Das et al., 2009) pour *Saccharomyces cerevisiae*. S'il est corrélé à la disponibilité en tRNAs, il permet d'augmenter l'efficacité et la justesse de la traduction (Michely et al., 2013; Qian et al.; Stoletzki and Eyre-Walker, 2007).

Le biais d'usage des codons chez les eucaryotes, comme chez les procaryotes, est très étudié dans le domaine de l'évolution moléculaire car il permet par exemple de repérer des transferts horizontaux ou de prédire l'expression des gènes mais aussi, de manière plus globale, pour étudier l'évolution puisque ce biais est le résultat de mutations sélectionnées et fixées dans une population.

De manière générale, lorsqu'une mutation par substitution est fixée, c'est à dire, lorsque les ancêtres communs les plus récents d'une population entière partagent cette mutation, on parle de substitution. La substitution d'un nucléotide par un autre au cours de l'évolution est un mécanisme basique dans l'évolution des séquences d'ADN. Ce processus est largement utilisé dans les études d'évolution moléculaire afin de déterminer le taux d'évolution et pour ainsi reconstruire l'histoire évolutive des organismes. Mais les changements dans les séquences ne sont pas directement observables car le phénomène de substitution est extrêmement long. Il est alors nécessaire de comparer au moins deux séquences descendant d'un même ancêtre commun.

1.2 L'étude de la sélection

1.2.1 Le nombre de substitutions entre deux séquences codantes

Pour estimer la divergence nucléotidique dans des régions non codantes, l'hypothèse posée est que tous les sites de la séquence ont le même taux d'évolution et qu'ils suivent le même patron de substitution. Mais comme nous avons vu plus haut, dans les régions codantes ce n'est pas le cas.

En effet quand les séquences comparées codent pour une protéine il est complexe de calculer le nombre de substitutions car il est nécessaire de distinguer les substitutions synonymes des substitutions non synonymes.

S'intéresser à cette divergence nucléotidique des régions codantes revient à se poser la question de la proportion des différences observées qui ont été fixées parce qu'elles sont bénéfiques ou parce qu'elle permettent une adaptation à l'environnement par rapport aux différences simplement fixées par la dérive génétique aléatoire (Bierne and Eyre-Walker, 2003).

La méthode qui semble être la plus complète pour évaluer la sélection est le test de MacDonald Kreitman (test MK) (McDonald and Kreitman, 1991) qui consiste à comparer le polymorphisme avec la divergence. Si le patron de mutations par substitution, c'est à dire le polymorphisme, diffère du patron de fixation, c'est à dire la divergence, alors il y a bien un effet de la sélection positive. Ce test est alors très robuste car il compare, dans les sites synonymes et non synonymes, les niveaux de variation intraspécifique avec la divergence interspécifique. Ceci permet d'estimer la proportion de substitutions fixées par la sélection positive plutôt que par la dérive génétique. Bien que cette méthode soit la seule permettant d'estimer sans biais le niveau d'évolution par sélection positive, elle demande beaucoup trop de données pour être appliquée dans notre cas.

Une autre méthode pour évaluer la sélection est alors l'estimation du taux des substitutions non synonymes sur les substitutions synonymes. Cette estimation se base sur le fait que s'il y a évolution adaptative sur la séquence, son taux d'évolution sera supérieur à celui d'une séquence neutre. Il est alors considéré que les mutations synonymes sont neutres et que les non synonymes sont soit délétères, soit neutres, soit avantageuses. Dans ce contexte, les gènes sont sous évolution adaptative si le taux de substitutions non synonymes (dN) est supérieur au taux de substitutions synonymes (dS) car les mutations avantageuses ont une plus grande probabilité de se propager dans une population que les mutations neutres. Comme la plupart des mutations non synonymes seraient délétères, dN tend à être largement inférieur à dS et dN est supérieur à dS seulement s'il y a effectivement évolution adaptative.

1.2.2 Les méthodes de calcul du dN/dS

Pour déterminer le nombre de substitutions synonymes et non synonymes il faut d'abord des informations sur le degré de sélection agissant sur la séquence.

La première étape consiste à compter les sites synonymes (S) et non synonymes (N) pour chaque codon sachant que chaque site peut être partiellement à la fois synonyme et non synonyme. Ensuite il faut déterminer le nombre de différences synonymes et non synonymes entre les séquences. Enfin, il est nécessaire d'appliquer une correction pour les substitutions multiples en un même site pour obtenir le dN et dS entre deux séquences. Bien que cette méthode apparaisse simple, il ne faut pas oublier que, comme expliqué précédemment, les séquences sont soumises à des biais différents de transversion, de nucléotides et de codons. Par exemple si on considère un codon dégénéré deux fois, faut-il considérer la troisième position comme un site synonyme, comme un tiers de site synonyme ou bien comme une autre fraction de site synonyme dépendant du ratio transition/transversion et du niveau du biais des codons (Bierne and Eyre-Walker, 2003) ?

Il existe alors une multitude de méthodes pour calculer le dN/dS, chacune prenant plus ou moins en compte ces paramètres ce qui donne lieu à des biais plus ou moins grands en fonction des données étudiées. Pendant longtemps les deux méthodes les plus utilisées était celles de Nei et Gojobori NG (Nei and Gojobori, 1986) et de Li *et al.* LWL (Barth, 1985). Mais ces méthodes considèrent que la distribution des substitutions est aléatoire. Ainsi elles ne prennent pas en compte le biais de transition et sous estiment donc le nombre de sites synonymes, ce qui surévalue le taux de substitutions synonymes par site. La méthode de Ina (Ina, 1995) fut la première à prendre en compte ce biais dans toutes les étapes de l'estimation mais elle ne prend pas en compte les substitutions multiples en un même site.

Actuellement la méthode la plus utilisée et qui est une des plus complète est celle de maximum-likelihood (ML) développée par Goldman et Yang (Goldman and Yang, 1994). Ce modèle est différent des précédents car il considère le processus de substitution entre les codons et non plus entre nucléotides. Il est flexible et permet d'incorporer des facteurs biologiques tels que le biais de transition mais aussi le biais d'usage des codons.

1.2.3 Signification du dN/dS

De manière très simplifiée, un ratio supérieur à 1 implique une sélection positive, dite darwinienne, et un ratio inférieur à 1 implique une sélection purifiante, dite stabilisatrice, et enfin un ratio égal à un indique qu'il n'y a aucune sélection, on parle de sélection neutre. Mais l'interprétation des résultats est beaucoup moins tranchée. En effet une combinaison de sélection positive et de sélection purifiante sur différents sites dans le gène ou à des moments différents au cours de l'évolution peut s'annuler, donnant une valeur moyenne qui peut être

inférieure, égale ou supérieure à un. De plus chaque génome a des caractéristiques propres amenant à décaler la limite indiquant une sélection neutre et séparant la sélection purifiante de la sélection positive.

Après s'être intéressé, dans ce premier paragraphe, aux pressions de sélection qui agissent sur les gènes, nous allons maintenant nous pencher plus particulièrement sur les familles de gènes et sur les forces qui régissent leur évolution. De manière générale, l'évolution des gènes est facilitée lorsque les gènes appartiennent à une famille multigénique car une des copies peut évoluer indépendamment vers une nouvelle fonction. Kimura et Otha ont d'ailleurs émis, comme un des cinq principes qui régissent l'évolution moléculaire, que la duplication de gènes doit toujours précéder l'émergence d'un gène ayant une nouvelle fonction (Kimura and Ohta, 1974). La duplication de gènes, à la base de la création de familles multigéniques, et la sélection sont alors de puissants mécanismes évolutifs car elles entrainent une diversification et/ou une perte différentielle de gènes.

2 La création d'une famille de gènes

Les familles de gènes sont composées de groupes de gènes très similaires, reflétant l'hypothèse que tous proviendraient d'un même ancêtre commun. Les familles de gènes sont en effet la conséquence directe du fait que la plus grande partie des nouveaux gènes proviennent de duplication de tout ou partie du génome comprenant des gènes entiers.
Les familles de gènes fournissent alors des informations sur la façon dont de nouveaux gènes apparaissent et se diversifient. De même, leur organisation peut fournir des indices sur la façon dont l'expansion des gènes dans certaines familles est régulée. Enfin, la présence de multiples copies de gènes apparentées peut permettre de mettre en évidence des forces évolutives non visibles par l'étude des gènes individuels.

2.1 Les duplications

Quelles soient à l'échelle du génome complet, d'une partie d'un chromosome, d'un gène complet ou bien seulement d'une suite de nucléotides, les duplications sont reconnues depuis longtemps comme étant un processus évolutif majeur d'innovation et de spéciation chez les eucaryotes (Ohno, 1970). Il existe plusieurs sortes de duplications : les duplications génomiques qui concernent les génomes dans leur intégralité, les duplications de chromosomes complets, les duplications segmentales qui touchent des régions génomiques et les répétitions satellites qui mettent en jeu seulement quelques nucléotides.

2.1.1 Les duplications satellites

Les duplications en tandem de domaine de quelques nucléotides sont appelées des répétitions satellites. Ces répétitions se distinguent par leur taille. Ainsi chez les levures la frontière entre les microsatellites et les minisatellites est fixée à 9 nucléotides (Richard and Dujon, 2006) et la longueur des megasatellites va de 135 à 417 nucléotides (Thierry et al., 2008). Beaucoup de protéines contiennent des satellites, plus particulièrement chez les eucaryotes (Bjorklund et al., 2006). Ce type de duplications concerne au moins 14% de toutes les protéines et correspond le plus souvent à des domaines structuraux ou fonctionnels en multicopies (Marcotte et al., 1999). Ceci suggère que beaucoup de gènes codants des protéines sont formés par des duplications internes et que les fonctions de ces protéines sont améliorées par l'augmentation de leur stabilité et de leur nombre de site actifs. De même, d'après Frenkel *et al.* les séquences évoluent en grande partie grâce à des évènements locaux de duplication de triplets suivies par des mutations dans les répétitions. Chez *E. coli* cette expansion des triplets serait la principale origine des gènes et ces répétitions serait statistiquement détectables dans un tiers des séquences modernes (Frenkel et al., 2013).

Bien que pour certains gènes et depuis peu pour le génome complet de *E. coli* (Frenkel et al., 2013), les duplications en tandem aient pu être mises en évidence, dans la majorité des cas les régions dupliquées ont tellement divergé l'une de l'autre que leur homologie n'est plus détectable. La reconstitution des évènements évolutifs et la détermination de la séquence ancestrale seraient possible seulement par comparaison des structures secondaires des protéines, lesquelles sont beaucoup mieux conservées que les acides aminées (Andrade et al., 2001). L'étude des duplications internes est très limitée par le peu de duplications encore observables dans les génomes et par la complexité de reconstruction des évènements évolutifs. Richard *et al.* ont montré que chez les levures hémiascomycètes même si de nombreux gènes contenant des minisatellites sont conservés, la plupart du temps le minisatellite lui-même n'est pas conservé (Richard and Dujon, 2006). Egalement, la diversité des séquences minisatellites trouvées dans les gènes orthologues de différentes espèces suggère que les minisatellites sont différentiellement acquis et perdus pendant l'évolution des levures hémiascomycètes à un rythme plus rapide que les gènes qui les contiennent (Richard and Dujon, 2006).

Les duplications satellites peuvent aboutir à des innovations dans les génomes mais ne permettent pas la création de nouveaux gènes et donc de familles multigéniques. Elles ne seront donc pas traitées dans cette étude.

2.1.2 La duplication complète du génome

La forme la plus remarquable de duplication est la duplication complète du génome ou polyploïdisation. La polyploïdie est un processus qui s'est produit à plusieurs reprises dans le règne animal et qui est très commun chez les plantes et plus particulièrement chez les angiospermes (pour revue voir (Otto and Whitton, 2000)). On parle d'autopolyploïdie lorsque la duplication des chromosomes se fait au sein de la même espèce. Par contre lorsqu'il y a hybridation entre deux ou plusieurs espèces, on parle d'allopolyploïdie. La polyploïdie peut être à l'origine d'une spéciation. Maclean *et al.* montrent d'ailleurs qu'après une duplication complète du génome induite expérimentalement chez un hybride SACE/*S. uvarum*, la perte réciproque des copies de gènes conduit à l'isolement reproductif partiel (Figure 3.1) (Maclean and Greig, 2011).

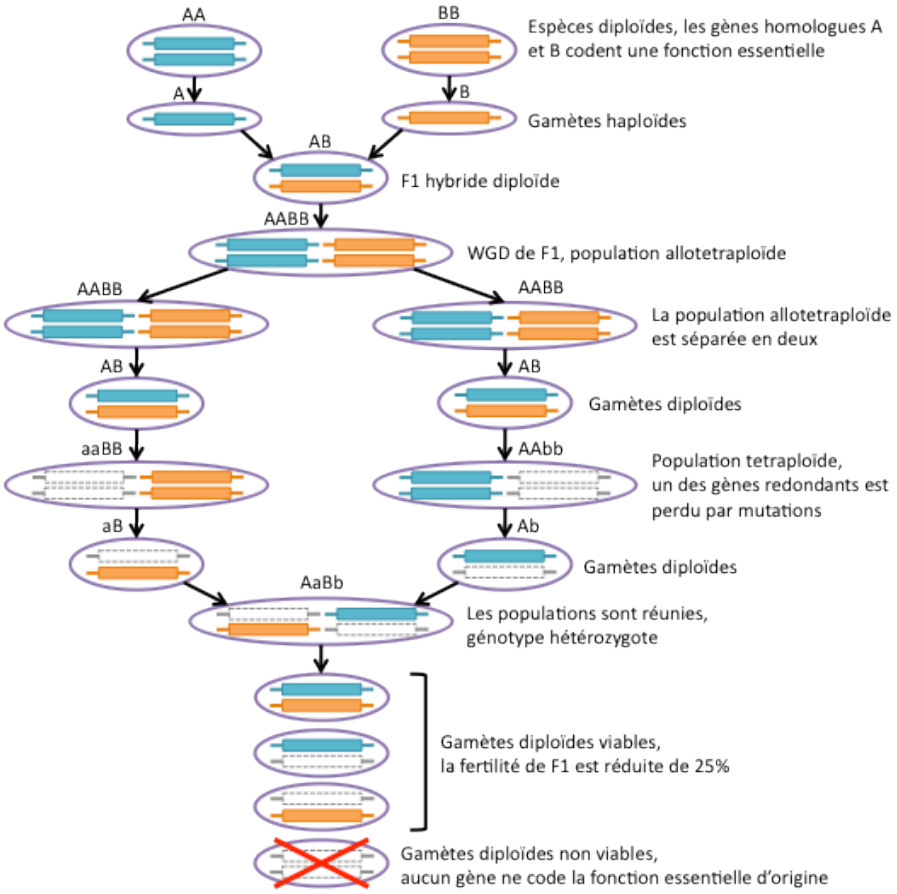


Figure 3.1 : Schéma montrant comment la perte réciproque d'un gène essentiel dans une population allotétraploïde peut entrainer l'isolement reproductif postzygotique d'après Maclean *et al.*. La population allotétraploïde provient de l'hybridation de deux espèces et de la duplication du génome. Les génotypes sont marqués au dessus des cellules. Les allèles A, B et a et/ou b sont respectivement représentés par des rectangles bleus, oranges et blancs. (Maclean and Greig, 2011)

SACE résulte elle même d'une duplication complète du génome WGD ayant eu lieu chez les *Saccharomyces* juste après la divergence *K. waltii* (Kellis et al., 2004; Wolfe and Shields, 1997). La WGD s'est produite soit par autopolyploïdie soit par allopolyploïdie, c'est à dire par la fusion de deux proches ancêtres diploïdes, chacun contenant environ 5000 gènes. Le tétraploïde est ensuite devenue diploïde et a subi une perte massive de gènes, environ 85% des copies dupliquées, conduisant à l'espèce actuelle SACE contenant ~5800 gènes et aux autres *Saccharomyces* (Figure 3.2). Les paires de gènes ohnologues, c'est à dires les gènes homologues provenant de la WGD, sont au nombre de 547 dans le génome de SACE (Byrne and Wolfe, 2005). Au niveau métabolique, la WGD conduit à une meilleure adaptation à l'environnement que les duplications à petite ou moyenne échelle. En effet la duplication d'une voie métabolique dans son intégralité, plutôt que de seulement une réaction, conduit à l'amélioration du phénotype. Chez SACE il a été montré que la WGD a joué un rôle essentielle dans l'adaptation au nouvel environnement, riche en glucose, qui a résulté de l'apparition des angiospermes (Gavin and Kenneth, 2007; Merico et al., 2007). Même si la WGD est très efficace pour l'adaptation à un nouvel environnement, elle est souvent préjudiciable dans un environnement pour lequel l'organisme est déjà adapté. Ceci explique pourquoi, malgré les importantes innovations qu'elle apporte et son apparente facilité à opérer, la WGD reste un événement rare (van Hoek and Hogeweg, 2009). Dans le cas où le milieu ne change pas ou peu, les duplications de un ou plusieurs gènes restent plus efficaces.

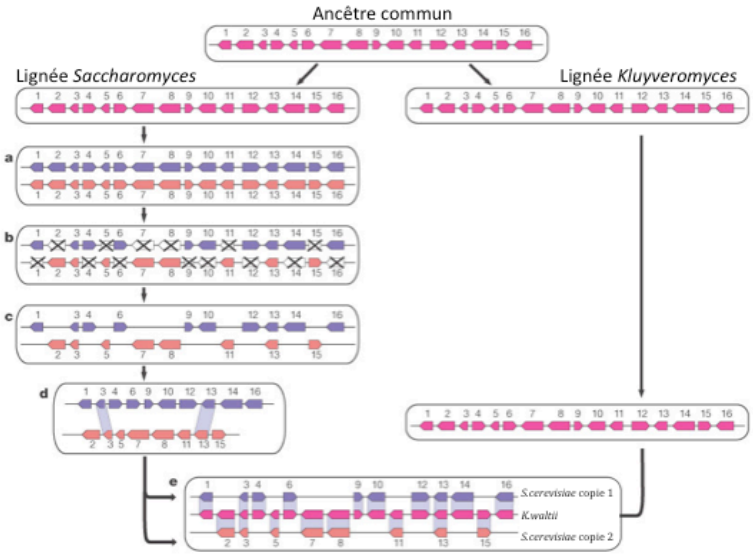


Figure 3.2 : Figure de Kellis *et al.*. **a)** Après la divergence de *K. waltii*, la lignée *Saccharomyces* a subi un évènement de duplication du génome, chaque chromosomes et chaque gène sont dupliqués. **b)** La majorité des gènes dupliqués ont subi des mutations et sont perdus. **c)** Les segments qui étaient homologues ont conservés différents sous ensembles des gènes d'origine. Quelques rares gènes dupliqués sont conservés en copie à des fins fonctionnelles. **d)** Dans *S. cerevisiae*, l'ordre conservé des gènes dupliqués (ici numérotés 3 et 13) à travers différents segments chromosomiques est une preuve de la duplication du génome. **e)** La comparaison avec *K. waltii* révèle un entrelacement de segments de gènes homologues par rapport à l'ordre ancestral montrant bien la nature dupliquée du génome de *S. cerevisiae* (Kellis et al., 2004).

2.1.3 L'aneuploïdie

Il existe aussi des duplications affectant un ou plusieurs chromosomes mais pas la totalité du génome, c'est l'aneuploïdie. L'aneuploïdie correspond à un nombre non homogène de copies de chromosome dans une même cellule qui peut être par défaut s'il manque un ou plusieurs chromosome(s) ou par excès s'il y a gain de chromosome(s). L'aneuploïdie est fréquente chez les plantes mais elle est très mal supportée dans le règne animal. Chez les mammifères une grande partie des maladies génétiques sont dues à ce phénomène, nous pouvons citer par exemple chez l'homme la trisomie 21 qui est l'excès d'un chromosome 21. Par contre, l'aneuploïdie par excès semble être très répandue chez *S. cerevisiae*. Hughes *et al.* ont montré que des souches de *S. cerevisiae* avec un chromosome supplémentaire peuvent avoir une meilleure croissance que les souches sauvages et que l'aneuploïdie par excès permet de complémenter les mutations délétères. Ils suggèrent alors que la prévalence des duplications de chromosomes dans la nature est peut être le résultat de la nécessité de compenser la perte de fonction des gènes au cours de l'évolution (Hughes et al., 2000).

2.1.4 Les duplications segmentales

Ces deux derniers types de duplications, WGD et aneuploïdie, sont des évènements remarquables dans l'évolution car ils changent drastiquement la quantité de matériel génétique. D'autres sortes de duplications, les duplications segmentales, ne touchent que quelques kilobases. Elles correspondent aux duplications d'une région génomique ou de tout ou partie d'un gène. Il y a alors apparition de deux copies ou plus, orientées dans le même sens ou inversées, localisées sur le même chromosome ou n'importe où ailleurs dans le génome (Koszul and Fischer, 2009).

2.1.5 Les duplications géniques : moteur de l'évolution

Il existe un cas particulier de duplications segmentales, les duplications géniques, qui ne concernent qu'un seul gène mais dans son intégralité. Du point de vue de l'évolution, les évènements de duplication/délétion s'appliquant aux gènes entiers sont très intéressants à étudier. Ohno et Lynch ont présenté la duplication génique comme un puissant moteur des changements évolutifs (Lynch and Conery, 2000; Ohno, 1970) et Olson introduit la perte de gènes comme une réponse avantageuse à des changements environnementaux (Olson, 1999). Par exemple, *S. pombe* et SACE partagent 56 familles de gènes dans lesquelles une ou plusieurs duplications se sont produites indépendamment dans chaque espèce après leur dernier ancêtre commun. Parmi ces 56 familles une grande partie ont un rôle crucial dans le cycle cellulaire, suggérant que les duplications dans ces familles sont impliquées dans l'adaptation de *S. pombe*

et SACE au cycle de vie unicellulaire de la levure (Hughes and Friedman, 2003). Nous pouvons également citer l'exemple de *Aspergillus fumigatus.* Chez ce champignon filamenteux ascomycète, 8,5% des gènes de son génome sont espèces spécifiques et sont souvent impliqués dans les métabolismes des carbohydrates ou des acides aminés, le transport, la détoxification ou des métabolismes secondaires. L'étude de ces gènes montre qu'ils proviennent de duplications suivies de diversification et/ou de perte différentielle et non pas de transferts horizontaux. Ceci suggère que la duplication est bien le premier mécanisme génétique à l'origine des gènes espèce spécifiques, apportant ainsi une grande diversité (Fedorova et al., 2008).

Tous les génomes montrent en effet de nombreuses traces de duplications géniques grâce auxquelles ils ont acquis de nouvelles fonctions. La Duplicated Genes Database (http://dgd.genouest.org/) recense 3775 gènes dupliqués chez l'humain soit plus de 12% des gènes du génome complet. Les duplications géniques récentes dans les génomes eucaryotes sont approximativement estimées à 0,01 par gène par million d'années avec une variation interspécifique très grande pouvant aller de 0,03 pour *Encephalitozoon cuniculi* à 0,001 évènements pour *Drosophila melanogaster* (Lynch and Conery, 2000 ; Lynch and Conery, 2003a). Chez *Saccharomyces cerevisiae* et *Schizosaccharomyces pombe* elles sont de l'ordre de 0,004 / gène / million d'années (Lynch and Conery, 2003a). Mais ces estimations du taux de duplications sont basées sur l'hypothèse selon laquelle les mutations génétiques s'accumulent dans un génome à une vitesse globalement proportionnelle au temps géologique, appelée modèle de l'horloge moléculaire. Ces estimations ne prennent pas en compte le phénomène d'évolution concerté, c'est à dire l'évolution non indépendante du nombre de copies dans une famille multigénique. Ceci conduirait à surévaluer le taux de duplications par gènes par années. Gao *et al.* ont alors réévalué ce taux en comparant d'une part un modèle basé sur l'horloge moléculaire et d'autre par un modèle prenant en compte l'évolution concertée (Gao and Innan, 2004). Ils estiment alors que SACE aurait un taux de duplication extrêmement faible, de l'ordre de 0,01/gène/milliard d'années contre 2,3/gène/milliard d'années en utilisant le modèle de l'horloge moléculaire.

Les duplications/pertes géniques, en plus d'être de grands moteurs pour l'adaptation à l 'environnement, jouent également un grand rôle dans l'organisation des génomes. En effet, Fischer *et al.* ont montré que, chez SACE et *Saccharomyces uvarum*, l'ordre des gènes est le plus souvent dû à des mécanismes impliquant des duplications géniques (Fischer et al., 2001). En effet, après une duplication de un ou plusieurs gènes entre deux chromosomes chez l'ancêtre commun il y a perte de certaines copies dupliquées au sein des espèces actuelles. Les copies perdues n'étant pas les mêmes entre les deux espèces, il y a rupture de synténie dans ces régions. De même, après une duplication génique inversée chez l'ancêtre si les copies sont

différentiellement perdues entre les espèces actuelles alors les deux homologues apparaitront comme ayant une orientation inversée.

Chez YALI aucune trace de duplication complète du génome ou d'aneuploïdie n'a été mise en évidence. Par contre la redondance de son génome, c'est à dire le nombre de gènes par familles, est de 1,37. Ceci signifie que le génome a subi de multiples duplications géniques conduisant à la création de familles multigéniques. 60 familles multigéniques contenant de 3 à 26 gènes ont été recensées (Morris and Drouin, 2011). L'expansion de certaines familles impliquées dans le métabolisme lipidique est même remarquable. Les lipases par exemples sont au nombre de 16 chez YALI alors que SACE, par exemple, ne possède qu'une lipase potentielle (voir Chapitre 2, I, 1.1). De plus la redondance génique pour le métabolisme lipidique, égale à 1,94 (204 gènes / 105 familles), est supérieure à celle du génome complet. Ces éléments nous ont amené à nous intéresser particulièrement aux duplications de gènes conduisant à la création de familles multigéniques impliquées dans le métabolisme des lipides. Pour cela il faut tout d'abord s'intéresser aux mécanismes conduisant à la duplication génique, à la création de familles multigéniques et aux évolutions des copies de gènes pouvant conférer des avantages physiologiques.

2.2 Les mécanismes à l'origine de la duplication génique

S'il est clair que la majorité des duplications chromosomiques proviennent d'un dysfonctionnement au cours des divisions cellulaires, mitotiques ou méiotiques, les mécanismes impliqués dans les duplications segmentales restent plus controversés et aucun n'a été clairement établi. Cependant, on peut citer plusieurs types de mécanismes cellulaires susceptibles de conduire à la duplication d'un gène (Figure 3.4).

2.2.1 Les retroséquences

Certains éléments transposables codent une reverse transcriptase inverse qui, en plus de permettre la prolifération et les mouvements d'éléments transposable dans le génome, peut dans certains cas rétro-transcrire une matrice d'ARN messager ARNm (Figure 3.3 et 3.4A). Cet ADNc est ensuite inséré dans le génome par rétrotranscription ce qui génère des retroséquences correspondent à des copies ADNc de gènes fonctionnels mais n'ont ni les introns, ni le promoteur. Elles possèdent une queue poly(A) provenant du 3' de l'ARN messager après la transcription et sont flanquées de répétitions directes aux extrémités. Souvent, elles sont tronquées en 5'. Les retroséquences peuvent être des retrogènes fonctionnels ou des pseudogènes appelés retro-pseudogènes processés. De façon générale, l'ADNc issu de la transcription inverse d'un ARN a deux destinées possibles :

– soit il remplace par recombinaison homologue le gène dont il est issu. Dans ce cas, si le gène possédait un intron, celui ci apparaitra comme perdu ;

– soit il s'insère à une position chromosomique par recombinaison non homologue ou homologue avec des séquences Ty1 complètes ou LTR solo (Maxwell and Curcio, 2007).

Il existe différentes possibilités permettant ensuite la transcription d'un retrogène (Kaessmann et al., 2009). La première possibilité est que les retrocopies s'insèrent dans une séquence intronique d'un autre gène. Dans ce cas l'évolution ou la présence de signaux d'épissage peut permettre à la copie de s'intégrer dans de nouveaux variants d'épissage du gène hôte. Ces variants donnent lieu à une fusion de transcrits codants si la région codante de la rétrocopie est fusionnée à celle du gène de l'hôte ou non codante si la totalité de l'ORF provient de la rétrocopie. La seconde possibilité est que la rétrocopie s'insère dans une région transcrite avec la chromatine décondensée, augmentant ainsi l'accès de la machinerie transcriptionnelle au rétrogène. Une autre possibilité est d'utiliser un promoteur éloigné grâce à l'acquisition d'une nouvelle structure intron-exon en 5'UTR. La quatrième possibilité est d'utiliser un promoteur déjà existant, celui d'un rétrotransposon ou celui du gène parent si le rétrogène est inséré au bon endroit. Il peut aussi utiliser un promoteur quelconque, créant éventuellement une fusion avec l'ORF initialement contrôlée par ce promoteur. La dernière possibilité est une évolution *de novo* par substitutions de la région 5' flanquant le site d'insertion pour générer un promoteur.

Un grand nombre de rétrogènes fonctionnels a été découvert principalement chez les mammifères et la drosophile. Ces études ont montré qu'ils ont souvent acquis des rôles fonctionnels dans la lignée germinale mâle, mais on peut citer d'autres rôles tel que la défense antivirale, le métabolisme des phéromones, dans le cerveau… (Kaessmann et al., 2009). SACE possède très peu de retro-pseudogènes processés alors qu'elle possède bien une reverse transcriptase, codée par Ty. D'autre par elle possède très peu de gènes à introns. Fink *et al.* proposent alors que chez SACE la plupart des gènes sont en fait des rétrogènes qui auraient remplacé les gènes initiaux par recombinaison homologue entrainant une perte des introns (Fink, 1987). Ceci est appuyé par le fait que les processus de recombinaison homologue sont très efficaces dans cet organisme, ainsi le principal site cible pour une rétrocopie est l'emplacement du gène qui a produit le transcrit. La recombinaison de l'ADNc ne génèrerait donc que rarement des retro-pseudogènes ou retrogènes sur un nouveau site chromosomique. Il a également été montré que la formation de rétroséquences entraine des restructurations dans le génome de SACE. En effet il arrive souvent qu'une retroséquence fusionne avec la séquence d'un élément transposable Ty1. Cela donne naissance à des pseudogènes chimériques jouant un rôle important dans l'apparition des réarrangements chromosomiques (Maxwell and Curcio, 2007).

Les rétroséquences ont alors une influence importante sur les profils d'expression génique, sur la production de nouvelles fonctions de gène et sur l'organisation du génome.

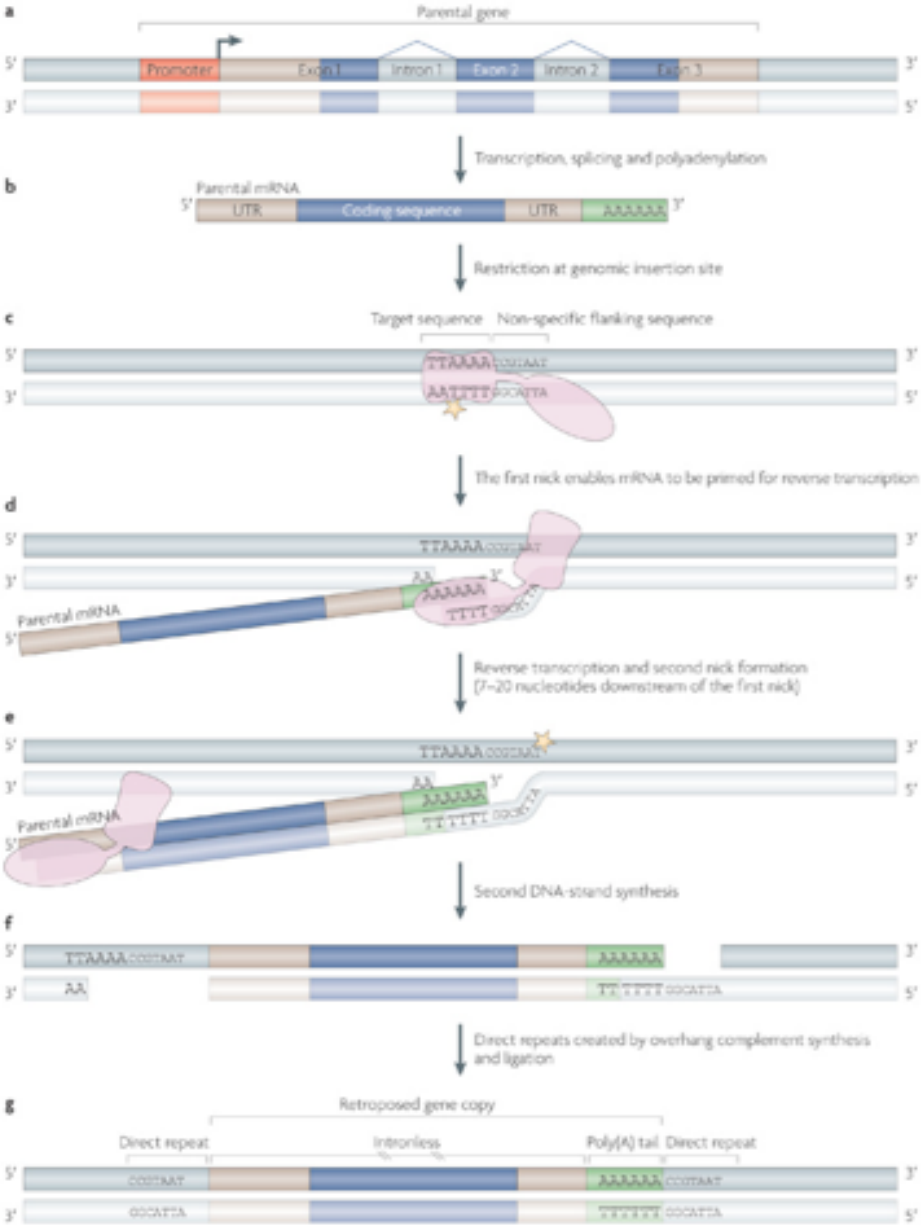


Figure 3.3 : Mécanisme de la rétroposition, de Kaessmann *et al.*. a) La rétroposition du gène est amorcée par la transcription d'un gène parental par l'ARN polymérase II. b) L'ARN obtenu est épissé et polyadénylé pour produire un ARNm mature. c) Une coupure d'un brin est créée (étoile jaune) à la séquence cible TTAAAA, futur site d'insertion, par le domaine endonucléase d'une reverse transcriptase (rectangle rose). d) Cette coupure permet d'amorcer la transcription inverse de l'ARNm parental par l'autre domaine de la reverse transcriptase (ovale rose). e) Le second brin est coupé. f) Le second brin d'ADN est synthétisé. g) Synthèse de l'ADNc dans les régions en excédant créées par les deux coupures. Ce processus crée une duplication de la séquence flanquant la séquence cible qui est une des signatures de la rétroposition génique en plus de l'absence d'intron et la présence de la queue poly-A. Par contre les répétitions directes et la queue poly(A) dégénèrent au cours du temps et ne sont donc en général détectables que dans les rétroséquences récentes. (Kaessmann et al., 2009)

2.2.2 La sur-réplication

Les duplications segmentales peuvent également provenir d'un événement de sur-réplication d'une région durant le processus de réplication (Figure 3.4B). Durant le cycle cellulaire, les phases S (duplication du génome) doivent alterner avec des mitoses et doivent reproduire le génome à l'identique. Pour dupliquer le génome il y a réplication de l'ADN à partir plusieurs origines. Grâce à différents mécanismes, chaque origine ne peut être initiée qu'une fois par phase S et seulement si elle à été précédée d'une mitose. Il arrive qu'une erreur se produise donnant lieu à deux types de sur-réplication : l'endoréplication et la re-réplication (Porter, 2008). Dans le cas de l'endoréplication plusieurs phases S se produisent consécutivement sans attendre la mitose. Dans ce cas l'origine de réplication n'est initiée qu'une seule fois par phase S. Durant la re-réplication l'origine peut être initiée plus d'une fois durant la phase S. Pour que le produit des fourches supplémentaires soit incorporé au génome il faut alors une cassure des fourches et un événement de recombinaison (Payen et al., 2008).

2.2.3 Recombinaison non homologue

Des évènements de réparation suite à des cassures non identiques de l'ADN sur deux chromosomes homologues peuvent également conduire à des duplications segmentales (Figure 3.4C). Dans ce cas il y a recombinaison entre deux séquences non homologues. Mais le mécanisme principal conduisant à une duplication de gènes est la recombinaison ectopique entre des séquences répétées en tandem. Lorsque la recombinaison ectopique a lieu entre deux chromosomes on parle de recombinaison inégale car dans ce cas l'un des chromosomes porte une duplication en tandem et l'autre une délétion (Figure 3.4D). Ces mécanismes peuvent conduire à la duplication d'un gène ou de plusieurs gènes adjacents, à l'aneuploïdie ou la polyploïdie. Chez SACE, dans le cas des ADN ribosomiques, les crossing-over inégaux entre chromatides sœurs sont fréquents (1% par cellule et par génération) et entrainent une variation du nombre de copies (Szostak and Wu, 1980).

Par contre les recombinaisons ectopiques entre des séquences répétées dispersées conduisent à des délétions ou des inversions quand les séquences sont situées sur un même chromosome et à des translocations réciproques ou à des chromosomes dicentriques ou acentriques quand elles sont sur des chromosomes différents.

La recombinaison ectopique entre des séquences répétées en tandem est largement facilitée par cette présence de séquences répétées. Ainsi à chaque fois que la séquence est dupliquée en tandem, le processus est accéléré car la probabilité d'un crossing-over inégal augmente avec le nombre de copies dupliquées (Figure 3.5).

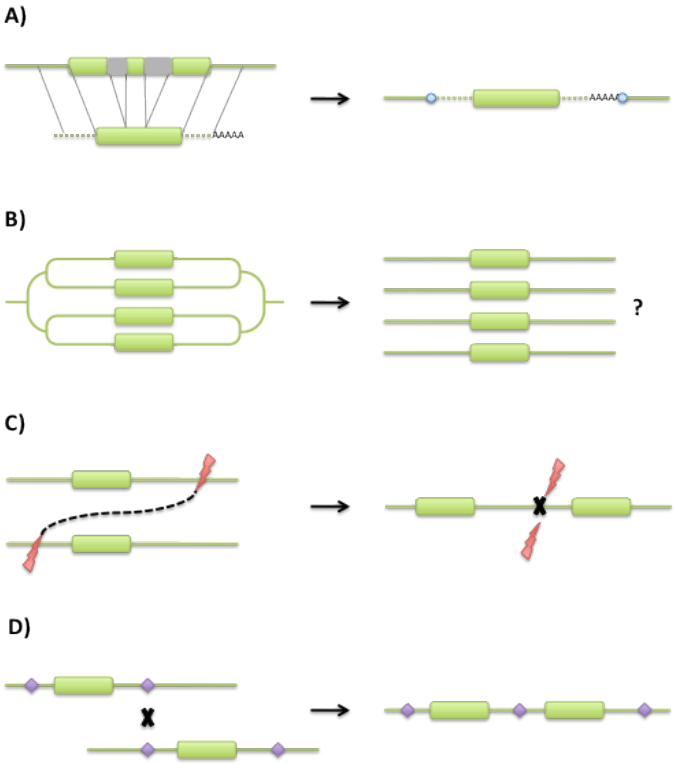


Figure 3.4 : Évènements pouvant conduire à la duplication de gènes d'après Maeda and Smithies (Maeda and Smithies, 1986). A) Rétrotranscription d'une matrice d'ARN messager (ARNm) en ADN complémentaire (ADNc) et insertion dans le génome. B) Sur-réplication d'une région par initiations multiples de la réplication au niveau d'une origine de réplication. C) Réparation suite à des cassures de l'ADN non identiques sur deux chromosomes homologues. D) Recombinaison ectopique entre des éléments répétés.

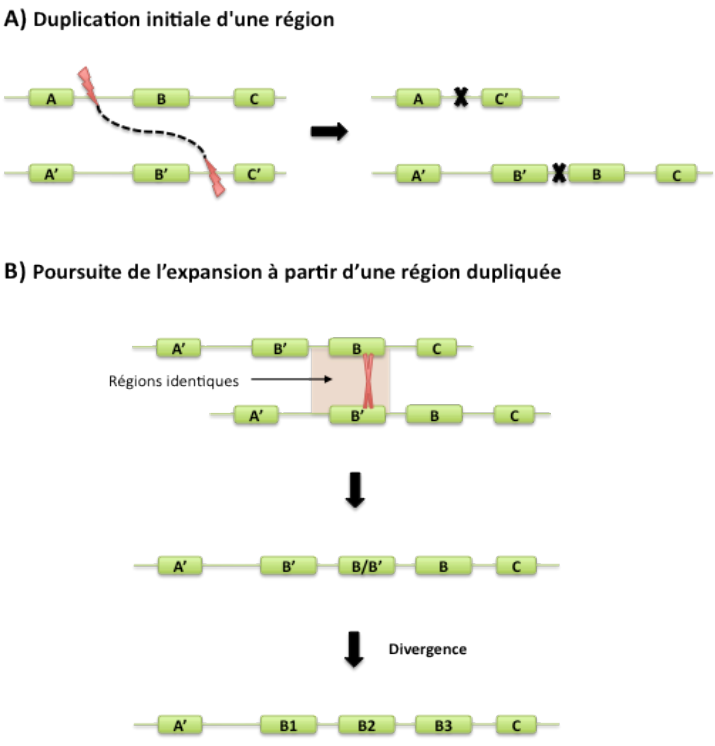


Figure 3.5 : Une famille de gènes générée par un crossing-over inégal : A) un crossing-over inégal et ses deux produits B et B'. Dans cet exemple, la région dupliquée correspond à un gène complet. B) second crossing-over inégal entre les deux copies du gène d'origine. Dans cet exemple le génome est homozygote. Au cours du temps les trois copies peuvent diverger en trois unités fonctionnelles distinctes (B1, B2 et B3) d'une même famille de gène. (Figure adaptée de http://www.informatics.jax.org/silver/)

3 Qu'est ce qu'une famille de gènes ?

3.1 Homologue, orthologue, paralogue ou ohnologues ?

En 2000, Fitch soulève le problème lié aux définitions de la terminologie utilisée pour décrire les diverses relations biologiques et propose des définitions claires afin d'harmoniser leur usage. Il définit l'homologie comme la relation entre toute paire de caractères qui descendent, généralement avec divergence, d'un caractère ancestral commun (Fitch, 2000). Les événements de duplication suscitent l'apparition de copies multiples à partir d'une seule séquence, soit au sein d'une même espèce, soit au sein des espèces distinctes dérivées de la spéciation. Ces séquences, dont la similarité résulte d'une séquence ancestrale commune, sont alors dites homologues. La relation d'homologie est donc transitive, ce qui implique que si A et B sont homologues et que B et C sont homologues alors A et C sont également homologues. Fitch définit également l'orthologie comme étant la relation entre deux caractères homologues dont l'ancêtre commun se trouve chez le cénancêtre des taxa, c'est à dire chez l'ancêtre commun le plus récent, à partir desquels les séquences ont été obtenues. Les gènes orthologues sont donc des séquences dont le dernier ancêtre commun précède immédiatement un événement de spéciation (Figure 3.6). L'orthologie est réciproque mais pas transitive. Cette définition ne requiert pas qu'il s'agisse d'une relation un-un, ainsi un gène peut avoir plusieurs orthologues dans une même espèce.
Dans le même sens, Fitch définit la paralogie comme la relation entre deux caractères émanant d'une duplication de gène pour ce caractère. Les paralogues sont alors des séquences dont le dernier ancêtre commun précède immédiatement un événement de duplication (Figure 3.6). Cette définition n'inclut pas de condition d'appartenance à un même génome ni d'indication sur l'âge de la duplication. Sonnhammer et Koonin proposent alors une subdivision des paralogues en outparalogues si l'événement de duplication a eu lieu avant la spéciation menant au dernier ancêtre commun, et en inparalogues si cet événement s'est produit après la spéciation (Sonnhammer and Koonin, 2002).

L'apparition de paralogues est le résultat de duplications. Dans le cas particulier de la duplication complète du génome on parle d'ohnologues. L'ohnologie qui est la relation entre les paires de gènes homologues provenant directement de la WGD. SACE possède 30 familles de paralogues, avec de 3 à 40 gènes par famille, contre 547 ohnologues (Byrne and Wolfe, 2005; Morris and Drouin, 2011). D'après Katjul *et al.,* chez les levures ayant subit la WGD les ohnologues et les paralogues auraient les mêmes caractéristiques malgré la disparité des mécanismes à leur origine. Chez SACE, ils montrent que les ohnologues et les non ohnologues

ont la même fréquence de duplications complètes ou partielles ; que très peu de copies sont structurellement différentes et donc qu'ils subissent tous deux une forte sélection purifiante et que leur localisation génomique est similaire (Katju et al., 2009). Cette étude est contradictoire avec celle de Morris *et al.* de 2011. Ces auteurs montrent que 8,4% des paralogues sont répartis sur les mêmes chromosomes contre 4% pour les ohnologues ; que la conversion génique tend à être plus fréquente entre des gènes situés sur un même chromosome et plus fréquente entre les paralogues qu'entre les ohnologues et que les ohnologues sont soumis à une plus forte contrainte sélective que les paralogues ce qui serait dû au fait que les ohnologues sont impliqués dans des processus essentiels contrairement aux paralogues (Morris and Drouin, 2011).

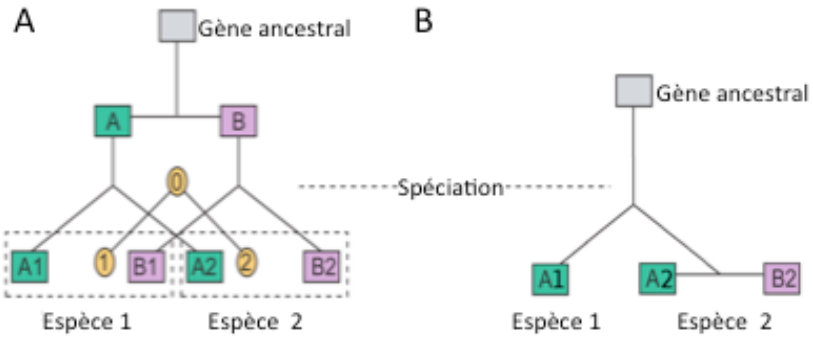


Figure 3.6 : Diagrammes simplifiés représentant des relations d'homologies tirés de Jensen *et al* A) Evolution des gènes paralogues A et B du génome ancestral 0. Après la spéciation créant les espèces 1 et 2, chacun des génomes contient deux copies du gène ancestral dupliqué (A1, A2 et B1, B2). Ces gènes sont tous homologues. Les couples de paralogues sont [A1, B1], [A1, B2], [A2, B1] et [A2, B2]. Les couples d'orthologues sont [A1, A2], [B1, B2]. B) Dans ce cas la spéciation a lieu avant la duplication. Le gène ancestral est dupliqué uniquement dans l'espèce 2. A2 et B2 sont orthologues de A1, [A2, B2] sont paralogues. (Jensen, 2001)

3.2 Les groupes d'orthologues et de paralogues

Tous les gènes appartenant à un groupe provenant d'une duplication ancestrale forment une famille multigénique et les protéines qu'ils codent sont homologues. Ces familles peuvent être constituées de paralogues, gènes issus d'un évènement de duplication, et d'orthologues, gènes homologues provenant d'une spéciation. Chaque paire de gènes du groupe présente alors une relation soit d'orthologie soit de paralogie par rapport au dernier événement de spéciation du clade. Dans ce contexte :

- les paralogues sont des gènes d'une même espèce, ils résultent d'une duplication ayant eu lieu après l'ensemble des événements de spéciation du clade ;
- deux gènes orthologues appartiennent à des espèces différentes.

4 L'ÉVOLUTION DES FAMILLES DE GÈNES

Plus d'une douzaine de modèles pour l'évolution et le maintien des gènes dupliqués ont été proposés ces dernières années (voir pour revue (Innan and Kondrashov, 2010)). Ces modèles diffèrent selon les forces de sélection et les évènements évolutifs ayant eu lieu au cours des différentes étapes que sont la duplication, la fixation dans la population, l'acquisition des différences et la préservation des nouvelles copies. Dans ce chapitre je ne traiterai que les modèles les plus communs.

4.1 Le processus de fixation neutre

La probabilité de fixation d'un nouvelle paire de gènes dupliqués A-A et la durée du la phase de fixation dépends de la force de sélection qui s'applique sur A-A et sur le génotype A. Si A-A est neutre, c'est à dire qu'il ne confère pas d'avantage sélectif ni de désavantage, il sera fixé dans la population à une fréquence de 1/2 N et le processus de fixation durera environ 4N (N étant la taille de la population). Ce type de modèle correspond à la néofonctionnalisation et à la subfonctionnalisation.

4.1.1 Néofonctionnalisation

Le premier modèle que nous pouvons citer est la néofonctionnalisation ou « mutation during non-functionality » MDN. Il correspondant à la création de nouvelles fonctions proposées par Ohno (Ohno, 1970). Plusieurs exemples montrent l'émergence d'une fonction entièrement nouvelle à partir d'une duplication. Un des exemples les plus cités reste celui de l'évolution de la protéine antigel du poisson Percidé de l'océan Antarctique. Cette protéine antigel est codée par un gène dérivé de sa copie paralogue codant l'acide salique synthase (SAS). La néofonctionnalisation nécessite généralement un grand nombre de substitutions, cependant dans certains cas une nouvelle fonction peut être obtenue avec très peu de substitutions. Par exemple la lactate déshydrogénase peut être changée en malate déshydrogénase en ne remplaçant qu'un de ces 317 acides aminés chez *Bacillus stearothermophilus* (Wilks et al., 1988). Dans le cas d'une néofonctionnalisation avec un grand nombre de changements génétiques nous pouvons penser que la protéine a probablement eu différentes fonctions au cours de ces changements. Selon le modèle MDN il a longtemps été admis que les gènes dupliqués étaient libres d'évoluer vers une nouvelle fonction tant que la fonction originale était maintenue par l'autre duplicat (Ohno, 1970). Cette prédiction dépend de l'hypothèse que l'une des copies reste sous sélection purifiante alors que l'autre, libre d'évoluer, est sous sélection neutre. Cependant trois forces s'opposent à cela. Premièrement, si la nouvelle fonction ne s'acquière pas

rapidement, après une ou deux substitutions, la probabilité de pseudogénisation devient beaucoup plus grande que celle de néo ou subfonctionnalisation. Deuxièmement, après la duplication les deux copies peuvent rester sous une pression purifiante aussi forte qu'avant la duplication. Et enfin, pour beaucoup de familles multigéniques divergentes, il a été mis en évidence que la divergence fonctionnelle des protéines est le résultat d'une sélection positive plutôt que neutre (Zhang et al., 1998).

4.1.2 Subfonctionnalisation

Cependant de nombreuses études ont montré que dans la plupart des cas il n'y a pas création d'une nouvelle fonction mais subdivision en deux ou plusieurs fonctions ancestrales entre les paralogues après la duplication. Dans ce cas on ne parle pas de néofonctionnalisation mais de subfonctionnalisation (Force et al., 1999). Dans chaque copie un domaine fonctionnel est rendu inactif par accumulation de mutations alors que l'autre domaine fonctionnel est conservé. Il y a alors partage des fonctions ancestrales entre les deux copies. La subfonctionnalisation peut correspondre au partage de l'expression génique, dans ce cas soit les copies sont exprimées dans des conditions différentes, soit les protéines sont adressées dans des compartiments cellulaires différents. Dans les deux cas les copies sont conservées dans le génome et se complémentent. La subfonctionnalisation est importante dans l'évolution car elle concerne de nombreux gènes.

SACE est un modèle de levure intéressant pour l'étude de la subfonctionnalisation et la néofonctionnalisation car la WGD subit par son ancêtre fournit un set de pairs d'ohnologues bien définis ayant le même âge et donc le même temps de divergence.
Marques *et al.* estiment que pour environ 30% des ohnologues les deux gènes d'une même paire codent des protéines ayant des adressages subcellulaires différents (Marques et al., 2008). L'évolution de la localisation des protéines limitées à un nombre restreint de compartiments est beaucoup moins fréquente que celle des protéines ayant une localisation cellulaire large. L'évolution du patron de localisation peut résulter du partage des fonctions ancestrales, subfonctionnalisation, mais il se réalise le plus fréquemment par relocalisation, c'est à dire par néofonctionnalisation. Les changements de localisation sont dus à une sélection agissant sur les séquences d'adressages. Ces protéines relocalisées montrent en plus des adaptations à leur nouvel environnement subcellulaire et ont souvent évolué vers de nouvelles fonctions grâce à des changements dans leurs propriétés physico-chimiques ou leur niveau d'expression par exemple.

Il a été montré, par exemple, que le caractère bifonctionnel de l'aminotransférase KlBAT1 de *K. lactis,* ascomycète dont les ancêtres n'ont pas subi la WGD, a été distribué dans deux gènes

ohnologues chez SACE, ScBAT1 et ScBAT2 (Colón et al., 2011). KlBAT1 participe à la fois à la biosynthèse et au catabolisme des acides aminés branchés dans la mitochondrie. Chez SACE une subfonctionnalisation à entrainée un partage de l'expression génique relatif aux conditions et à la localisation. En effet ScBAT1 est fortement exprimé en conditions biosynthétiques et en condition fermento-respiratoire et la protéine *ScBAT1* génère les précurseurs biosynthétiques dans les mitochondries alors que ScBAT2 est fortement exprimé en condition catabolique et en condition de respiration et *ScBAT2* génère les substrats pour le catabolisme dans le cytosol.
Nous pouvons également citer l'exemple la paire d'orthologues ScALT1 et ScALT2 qui partagent 65% d'identité. Chez *L. kluyveri*, dont aucun ancêtre n'a subit de WGD, l'activité alanine aminotransférase est codée par le seul gène LkALT1 dont l'activité est plus forte en présence d'alanine. Chez SACE, ScALT1 code une aminotransférase alors que ScALT2 code une protéine catalytique inerte. *ScALT1* est la seule voie possible pour le catabolisme de l'alanine et le gène est régulé par la présence de cet acide aminé alors que ScALT2 est réprimé par la présence d'alanine (Peñalosa-Ruiz et al., 2012). Cette répression pourrait indiquer un rôle dans sa biosynthèse mais *ScALT2* n'a pas d'activité aminotransférase. Peñalosa-Ruiz *et al.* proposent que ScALT2 aurait été retenu dans le génome car l'enzyme codée jouerait un rôle de modulateur d'autres produits génique intervenants dans la biosynthèse de l'alanine mais aussi parce qu'elle aurait un autre rôle physiologique en relation avec l'adaptation au stress environnementaux.

Actuellement deux principaux modèles de subfonctionnalisation sont proposés, la duplication-dégénération-complémentation DDC et la spécialisation, appelée plus récemment 'Escape from adaptative conflict' (Des Marais and Rausher, 2008). Ils différents selon la force de sélection qui s'exerce sur les gènes au cours de la phase de différenciation, après la duplication.

2.a La duplication-dégénération-complémentation DDC

Dans ce modèle, proposé par Force *et al.*, il y a accumulation de mutations neutres qui endommagent la fonction du gène, réduisant l'efficacité de chacune des deux copies (Force et al., 1999). Chacune des deux copies devient alors insuffisante pour assurer la fonction originale ce qui conduit au maintien des deux copies par la sélection. Dans ce modèle, durant la phase d'acquisition des différences les deux copies sont sous sélection purifiante mais de faible intensité. A la fin de cette phase, la fonction originale sera divisée en deux gènes en terme d'expression ou de fonction de la protéine de manière neutre sans sélection positive. Une fois que les deux copies ont acquis des fonctions différentes, le taux d'évolution de chaque copie peut augmenter afin d'améliorer leur fonction respective.

2.b Spécialisation

Dans ce modèle Hugues propose que la copie ancestrale possède déjà deux fonctions qui ne peuvent être améliorées indépendamment (Hughes, 1994). Ainsi après la duplication chaque copie sera soumise à la sélection positive afin de se spécialiser en améliorant l'une des deux fonctions ancestrales. Comme pour le modèle DDC, après la phase de différenciation, la divergence de chaque copie est soumise à sélection positive afin d'améliorer leur fonction respective.

Les modèles de DDC et spécialisation peuvent également aboutir à la pseudogénisation d'une des copies. En effet la sélection est relaxée durant la phase de différenciation donc les copies doivent acquérir une nouvelle fonction rapidement pour entrer dans la phase de préservation.

4.1.3 La pseudogénisation

Une des copies peut également subir une mutation délétère et devenir non fonctionnelle. Cette copie devient alors un pseudogène. En effet il est souvent peu ou pas avantageux d'avoir une redondance fonctionnelle. De plus les mutations agissant sur la fonction et la structure d'un gène présent en deux copies ne sont pas délétères pour l'organisme et donc ne sont pas contre sélectionnées. Au cours du temps les mutations s'accumulent dans le gène et celui-ci devient alors un pseudogène non exprimé et/ou sans fonction. Seuls les pseudogènes récents sont identifiables par similarité car au cours de l'évolution les pseudogènes sont soit éliminés du génome soit ils deviennent non reconnaissables tellement ils ont divergé depuis le gène ancestral. La pseudogénisation, processus par lequel un gène fonctionnel devient un pseudogène, a généralement lieu peu de temps après la duplication, lorsque le gène n'est pas sous sélection. Mais le gène dupliqué peut aussi avoir acquis une fonction spécialisée depuis longtemps avant sa pseudogénisation suite à un relâchement des contraintes fonctionnelles. Dans certains cas les pseudogènes peuvent même redevenir fonctionnels au cours de l'évolution. On peut citer l'exemple du gène *IRGM* membre de la famille des Immunity-related GTPases (*IRG*) qui a été pseudogénisé au cours de l'évolution humains-primates et dont l'ORF à été « ressuscitée » chez l'humain et dans la lignée des grands singes suite à une série d'évènements structuraux complexes (Bekpen et al., 2009).

Lafontaine *et al.* ont caractérisé les pseudogènes de 8 espèces de levures dont SACE et YALI. Les 6 autres espèces, *C. glabrata, K. thermotolerans, S. kluyveri, Z. rouxii, K. lactis* et *D. hansenii,* ont respectivement dans leur génome 0,7%, 1,2%, 1,3%, 2,1%, 2,2% et 2,8% de pseudogènes (Lafontaine and Dujon, 2010). Ces auteurs comptent 77 pseudogènes chez SACE (dont 44 entiers et 33 tronqués), soit 1,3% de son génome et 230 pseudogènes chez YALI (dont 82 entiers et 148

tronqués), soit 3,6% de son génome. Parmi les 230 pseudogènes de YALI, seulement 6 ne proviennent pas d'une duplication mais d'une simple perte de fonction d'un gène en simple copie. Par contre 224 proviennent de la dégradation d'un gène précédemment dupliqué, soit avant la divergence de l'espèce (11 cas) soit de manière espèce spécifique (213 cas). Parmi les pseudogènes ayant pour origine la duplication espèce spécifique d'un gène, 10 seraient potentiellement des retro-pseudogènes processés. Cet article montre également que YALI se distingue par la localisation de ses pseudogènes. Dans les 7 autres espèces étudiées, ils sont répartis en clusters dans les régions subtélomériques où ils sont 3 à 16 fois plus nombreux que les gènes fonctionnels. Chez YALI seulement 5% des pseudogènes sont dans les subtéloméres, contre 71% pour SACE. Harrison *et al.* recensent chez SACE 138 dORF définies comme des cadres de lecture ouverts (ORF) de pleine longueur pouvant coder des protéines mais contenant des invalidités évidentes tels que des sauts de phases ou des codons stop prématurés. Ils confirment que ces dORF sont les gènes les plus susceptibles de devenir des pseudogènes et de varier en nombre de copies. Ils confirment que leur proportion est très élevée dans les régions subtélomériques et mettent en évidence le rôle des régions à engendrer la diversité et l'évolution des protéomes de levures(Harrison et al., 2002).

NB : Avec la réannotation de YALI nous recensons seulement 141 pseudogènes. Ceci vient du fait que nous n'avons pas compté les séquences trop courtes.

4.2 Fixation par la sélection positive

4.2.1 Répétition de dose

Les copies peuvent garder la fonction originale ce qui va permettre à l'organisme de produire une plus grande quantité d'ARN ou de protéines. En effet la présence de gènes dupliqués peut être bénéfique simplement grâce à l'augmentation de la quantité de protéines ou d'ARN correspondant. On parle alors de répétitions de dose. Ces répétitions sont favorisées lorsqu'elles sont justifiées par un besoin métabolique pour une molécule donnée. On peut citer par exemple les gènes codant les ARN de transfert présents en de nombreuses copies dans la majorité des génomes.

Dans ce modèle il y a sélection positive pour la duplication, augmentant ainsi la probabilité et la vitesse de fixation de la duplication dans la population. Concernant la phase d'accumulation des différences, deux possibilités existent pour que les paralogues gardent la même fonction. Si les deux gènes sont fortement soumis à la conversion génique alors ils vont garder des séquences et des fonctions très similaires. Ce type d'évolution est appelé l'évolution concertée, c'est à dire que les séquences répétées à l'intérieur d'une même espèce divergent moins entre elles qu'avec les

séquences répétées de la même famille d'une espèce voisine (voir paragraphe 4.3). L'autre phénomène qui permet à deux paralogues de garder la même fonction est une forte sélection purifiante contre les mutations qui entraineraient la divergence des deux copies et de leur fonction.

Plusieurs études sur des familles de gènes ont montré que la sélection purifiante est beaucoup plus importante que la conversion génique pour le maintien de la fonction commune des paralogues. Ceci à été montré par exemple pour les gènes de la famille multigénique des ubiquitines chez 28 espèces de levures, plantes et animaux et pour les hydrophobines de classe 2 du genre Trichoderma/Hypocrea (Kubicek et al., 2008; Nei et al., 2000).

4.2.2 « Innovation, amplification, divergence » IAD

Dans les processus neutre décrit précédemment nous avons vu qu'une nouvelle copie doit accumuler de rares mutations afin d'acquérir une fonction qui sera sélectionnée. Mais ces rares mutations ne peuvent être obtenues que si la copie du gène reste suffisamment longtemps dans la population et à une fréquence allélique suffisante. Une autre solution serait le maintien de la copie supplémentaire par la sélection. Cependant, nous avons vu que cette sélection aboutirait à une répétition de dose, empêchant alors la copie de perdre son activité originale et de gagner une nouvelle fonction.

Récemment, un autre modèle de subfonctionnalisation à été proposé : « innovation, amplification, divergence » IAD (Bergthorsson et al., 2007; Näsvall et al., 2012). Selon ce modèle l'établissement d'une nouvelle fonction se fait via une activité secondaire du gène parental. Un changement environnemental peut imposer une sélection pour augmenter le niveau d'une activité secondaire du gène parental. L'augmentation de cette activité secondaire sera préférentiellement réalisée par duplication du gène original car les phénomènes de duplication sont 4 à 8 fois plus fréquents que les substitutions sélectionnées. L'amplification du gène parental va alors augmenter en fréquence dans la population et sera maintenu par sélection positive. Cette sélection agissant en continu pour l'amplification permet de résoudre les problèmes s'opposant à la théorie d'Ohno. Cette sélection va également favoriser l'augmentation du niveau de la nouvelle activité produite par chacune des copies. De plus, le nombre de copies supplémentaires augmente la probabilité qu'une mutation favorable soit fixée et ces mutations peuvent être associées à des processus de recombinaison entre les copies permettant d'obtenir des améliorations de certaines copies. Les copies améliorées peuvent être encore amplifiées alors que d'autres copies moins fonctionnelles, incluant le gène ancestral peuvent être perdues. Comme dans le modèle de subfonctionnalisation, la duplication agit sur un gène bifonctionnel et les mutations conduisent à des pertes de fonction supprimant une des activités dans chaque gène. Ces gènes sont alors maintenus par la sélection pour leur activité distincte. Mais les modèle classiques MDN et subfonctionnalisation prédisent que la séquence d'un nouveau gène

peut montrer qu'il existe des périodes durant lesquelles la sélection n'aurait pas agi. En revanche le modèle IAD prévoit que les gènes avec de nouvelles fonctions soient sous sélection positive continue car elle agit à toutes les étapes du processus. De plus, les périodes sans sélection sont minimisées par le fait que les mutations spécialisantes sont communes à celles qui altèrent une des fonctions préexistantes plutôt qu'elles ne créent une nouvelle fonction.

4.3 Evolution divergente, concertée et « birth-and-death »

Les gènes composant une famille peuvent diverger graduellement pour obtenir de nouvelles fonctions. Dans ce cas les gènes évoluent de façon indépendante (Figure 3.8A). Mais ce modèle n'explique pas les nombreux cas observés dans lesquels les séquences des gènes sont plus similaires dans une espèce qu'entre deux espèces proches. Un autre modèle prédit alors que les gènes évoluent de façon concertée (Figure 3.8B).

L'évolution concertée est la conséquence des évènements de recombinaison, tels que la conversion génique, biaisée ou non, et le crossing over inégal, entre les gènes d'une même famille présentant une similarité de séquence. Suite à ces nombreux échanges de séquences, certaines familles de gènes semblent évoluer de concert, avec les membres d'une même espèce. Les gènes d'une famille sont alors plus semblables les uns aux autres que ce qu'ils ne le sont avec les gènes d'espèces étroitement apparentées (Zimmer et al., 1980). Kominek *et al.* ont récemment mis en évidence l'évolution concertée au sein de la sous famille des SSB, composée de deux paralogues, appartenant à la familles des chaperonnes HSP70 pour les génomes des espèces Saccharomycetaceae apparues après la duplication complète du génome (Figure 3.7). Ils présentent un arbre phylogénétique caractéristique de ce que l'on peut observer en cas d'évolution concertée puisque les paralogues se regroupe ensemble.

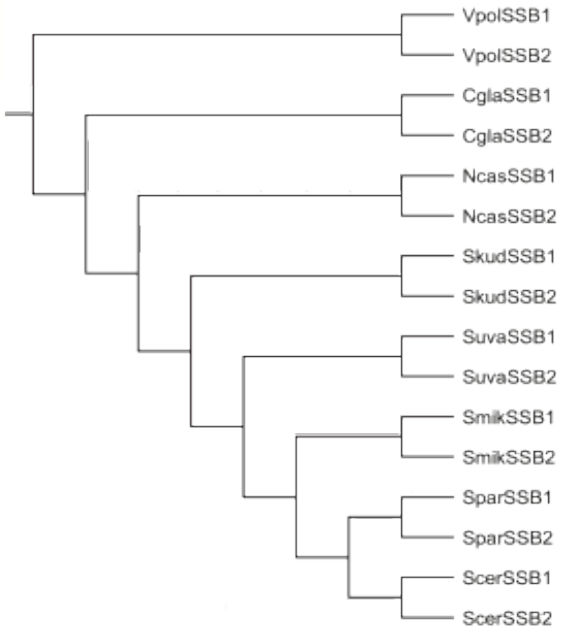


Figure 3.7 : Arbre phylogénétique des séquences nucélotidiques des SSB des espèces Saccharomycetaceae apparues après la duplication complète du génome. Vpol : *Vanderwaltozyma polyspora*, Ggla : *Candida glabrata*, Ncas: *Naumovozyma castellii*, Skud : *Saccharomyces kudriavzevii*, Suva : *Saccharomyces uvarum*, Smik: *Saccharomyces mikatae*, Spar : *Saccharomyces paradoxus*, Scer : *Saccharomyces cerevisiae*.

Bien que l'évolution concertée ait peu d'effet bénéfique dans des groupes de gènes fonctionnellement distincts, elle facilite la sélection agissant pour le maintien de la fonctionnalité dans une famille de gènes constituée d'un grand nombre de gènes fonctionnellement identiques. Lorsque la sélection agit de façon indépendante sur chaque membre, on s'attend à ce qu'un certain nombre de copies deviennent inactives cependant, les familles contiennent souvent un pourcentage élevé de copies fonctionnelles, ce qui est probablement le résultat d'une interaction entre l'évolution concertée et la sélection (Walsh and Stephan, 2001). Les cycles d'échange de séquence permettent en effet soit d'amplifier une mutation rare en la répartissant entre différents membres de la famille soit de la supprimer entièrement. Si la nouvelle mutation est délétère, l'augmentation de sa fréquence permet à la sélection d'agir plus efficacement pour la retirer. Si par contre la mutation est avantageuse, l'augmentation de sa fréquence permettra cette fois à la sélection d'agir plus efficacement pour sa diffusion dans toute une famille. Cette capacité à disperser une mutation entre tous les membres d'une famille a de grandes conséquences dans l'évolution : cela permet de disperser une mutation avantageuse, de retarder la divergence des paralogues, elle réduit la différence génétique entre les réplicats et augmente le niveau de variations génique à un locus donné (Nagylaki, 1984). L'évolution concertée complique la datation des évènements de duplication de gènes. En effet quand les séquences de paralogues sont vraiment très similaires il est impossible de distinguer entre les deux possibilités : soit les séquences commencent à diverger suite à une duplication récente, soit les séquences ont évolué de concert.

Cependant, des études plus récentes ont montré que des familles de gènes que l'on croyait évoluer de façon concertée évoluent en fait selon un processus d'évolution appelé « birth and death » (Rooney and Ward, 2005). Dans ce modèle, de nouveaux gènes sont créés par duplication et certains d'entre eux sont maintenus durant une longue période dans le génome alors que les autres sont supprimés ou deviennent inactifs par accumulation de mutations délétères (Figure 3.8C). Contrairement au modèle d'évolution concertée, c'est la sélection, et non l'homogénéisation, qui maintient les copies.

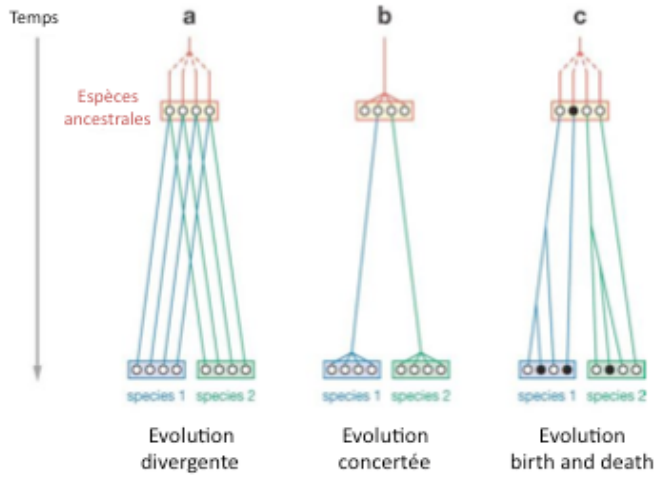


Figure 3.8 : Trois modèles différents d'évolution des familles multigéniques. Les cercle ouverts représentent les gènes fonctionnels et les cercles pleins représentent les pseudogènes (Nei and Rooney, 2005).

5 Retracer l'histoire évolutive des familles de gènes

Pour retracer le scénario évolutif d'une famille de gènes il faut tout d'abord déterminer les gènes qui constituent cette famille. Les paralogues, descendant d'une duplication, sont retrouvés par homologie. Par contre les orthologues, descendant d'une spéciation sont plus difficiles à retrouver et nécessitent d'utiliser plusieurs méthodes. Déterminer les groupes d'orthologues s'avère très intéressant car ces gènes, dont l'histoire évolutive correspond par définition à celle des espèces, ont généralement la même fonction et ils peuvent être utilisés pour prédire la fonction d'un gène, pour son annotation par exemple, en se basant sur les fonctions connues des autres orthologues. Ils peuvent aussi être utilisés pour inférer l'arbre phylogénétique des espèces, dans le cadre de la phylogénomique, et pour établir une correspondance un-à-un entre les gènes d'une paire de génomes (Guillaume, 2007). Ensuite, les familles de gènes servent à estimer les taux de duplications et de pertes de gènes et établir le contenu en gènes des génomes ancestraux (De Bie et al., 2006; Hahn et al., 2007; Tesler et al., 2007).

5.1 Identification des orthologues par homologie et synténie

Comme nous l'avons vu précédemment, les différences entre les séquences de deux gènes homologues résultent de l'accumulation de mutations à partir de l'ancêtre commun. Il s'agit donc d'une évolution par divergence évolutive. Pour affirmer que deux séquences sont homologues, l'idéal serait de pouvoir retracer leur histoire jusqu'à leur ancêtre commun mais il n'est malheureusement pas possible de disposer des séquences de tous les ancêtres disparus. On s'attend tout de même à ce que les gènes orthologues de deux génomes aient la similarité la plus élevée entre eux du fait de leur divergence relativement récente. Il est donc possible de repérer les orthologues en comparant simplement les gènes entre eux. Les gènes ayant les similarités les plus élevées sont alors considérés comme orthologues. La stratégie de "Bi-directional Best BLAST hit" est souvent utilisée afin de ne garder que les homologies les plus significatives. Le principe de cette stratégie est le suivant : soit gA un gène dans le génome A et gB un gène dans le génome B ; l'homologie de gA et gB n'est inférée que si gB est la meilleure occurrence trouvée dans B par une recherche par BLAST avec la cible gA, et inversement, gA est la meilleure occurrence trouvée dans A par BLAST avec la cible gB.

Mais, à de relativement grandes distances évolutives, la similarité entre des séquences orthologues devient comparable à celle observée entre des séquences appartenant simplement à la même famille de gènes. La similarité devient alors insuffisante pour poser l'hypothèse d'orthologie. Dans ce cas l'étude de la synténie est un outil puissant pour mesurer l'orthologie.

Par exemple, l'observation de la similarité des gènes contigus gA1 et gA2 du génome A avec les gènes gB1 et gB2 également contigus sur le génome B permet de conclure à leur orthologie. Par contre l'étude de la synténie n'est possible que sur des génomes dont la divergence est relativement récente, si ce n'est pas le cas l'ordre des gènes dans les génomes ne sera pas suffisamment conservé.

Cependant, l'étude conjointe de la similarité et de la synténie ne permet pas de supprimer le problème qui se pose dans le cas de la convergence évolutive. En effet, si deux gènes homologues qui n'ont pas de lien phylogénétique ont la même fonction dans deux organismes, alors même l'observation d'absence de synténie ne permettra d'affirmer qu'ils ne sont pas orthologues puisque rien ne les distingue d'un couple d'orthologues dont un des gènes s'est déplacé sur le génome.

5.2 La phylogénie d'une famille de gène

Pour déterminer la relation d'orthologie entre deux gènes de deux génomes différents appartenant chacun à une famille mutigénique, nous pouvons utiliser la synténie quant elle est conservée. Mais la plupart du temps il est nécessaire d'utiliser des méthodes phylogénétiques. Dans ce cas, la reconstruction de l'histoire évolutive de la famille de gènes peut aider à distinguer les relations d'orthologie et même de paralogie. Après un alignement multiple des séquences d'une famille de gènes homologues, l'arbre phylogénétique peut être reconstruit par des méthodes de parcimonie ou de maximum de vraisemblance. Construire une phylogénie selon le principe de parcimonie revient en fait à déterminer l'hypothèse qui maximise le nombre d'hypothèses d'homologie (similitude due à une ascendance commune) tout en minimisant le nombre d'hypothèses d'homoplasie (similitude due à une adaptation convergente) (Fitch, 1971). L'arbre phylogénétique d'une famille peut alors permettre de déterminer des groupes monophylétiques regroupant les orthologues. De plus, une phylogénie permet aussi de reconstruire le scénario évolutif d'une famille de gènes d'espèces relativement proches. Il est en effet possible de retracer une partie des évènements ayant eu lieu au cours de l'évolution de la famille, tel que les pertes ou les duplications de gènes, et de déterminer le nombre de gènes de cette famille chez l'ancêtre commun (AC). Une expansion ou *a contrario* une contraction du nombre de gènes dans la famille depuis la divergence de l'AC permet également de soulever un mécanisme d'évolution pouvant conduire à une adaptation dans les espèces. Une manière de reconstruire ces scénarios est d'utiliser le principe de parcimonie, postulant que l'évolution la plus vraisemblable est celle qui nécessite le plus petit nombre de changements évolutifs, ce qui revient à construire les scénarios évolutifs qui minimisent la somme des coûts sur toutes les branches. La construction de la phylogénie des familles de gènes est donc de façon générale un outil important pour l'étude de groupes d'orthologues et de paralogues.

5.3 Incongruité entre arbre de gènes et arbre d'espèces ou Arbres consensus

Les gènes appartenant aux espèces actuelles ont aussi leur propre histoire, qui peut être différente de celle des espèces, à partir desquelles des relations entre les gènes sont définies (Fitch, 2000). Il est effectivement courant que l'arbre des gènes appartenant à une famille ne soit pas isomorphe à celui des espèces. Dans ce cas, si l'on considère que les deux phylogénies sont robustes, l'incongruence entre un arbre de gènes et un arbre d 'espèces peut s'expliquer par l'évolution des gènes par duplication et pertes. Par exemple, l'incongruence entre l'arbre des gènes G et l'arbre des espèces, présentée dans la figure 3.9, peut s'expliquer par une duplication du gène G chez l'AC suivie par 3 pertes, une par espèce, du gène G.

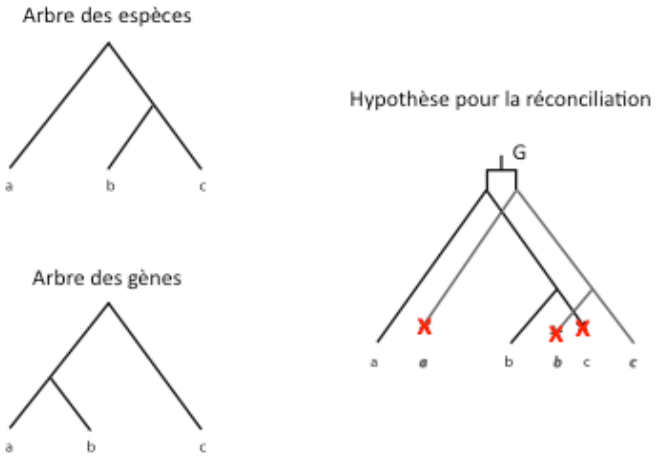


Figure 3.9 : Exemple d'incongruence entre un arbre de gènes et un arbre des espèces a, b, c. Une hypothèse possible est la duplication du gène G chez l'AC puis la perte d'une des copies dans chacune des espèces a, b et c. D'après (Doroftei, 2012).

Lorsque les topologies de l'arbre des gènes et l'arbre des espèces ne sont pas superposables il est donc possible de trouver une topologie commune aux deux arbres, c'est-à-dire l'arbre consensus. Goodman *et al.* sont les premiers à décrire et à utiliser le concept de réconciliation pour résoudre l'incongruence entre la phylogénie de vertébrés basée sur des données morphologiques et l'arbre de gènes des globines (Goodman et al., 1979). L'objectif d'une réconciliation est de trouver l'arbre de gènes G le plus similaire à la phylogénie d'espèces S pour lequel l'histoire évolutive de G selon S est parcimonieuse, c'est à dire qu'elle cumule le moins d'évènements de type mutation, perte ou duplication de gènes (Figure 3.10). Un arbre réconcilié nous permet de reconstruire de manière précise une histoire évolutive de gènes qui respecte la phylogénie des espèces. Malheureusement, quelle que soit la méthode d'inférence de phylogénie utilisée, de mauvaises constructions peuvent être obtenues. Ceci a des répercussions immédiates sur la cohérence de l'histoire évolutive obtenue par réconciliation entre un arbre d'espèces et un arbre de gènes. Effectivement, une branche mal placée dans l'arbre des gènes peut entraîner une réconciliation complètement différente. Les erreurs dans l'inférence des phylogénies peuvent être dues à un mauvais choix de méthode de construction d'arbre ou bien à un mauvais échantillonnage. Les échantillonnages disponibles sont souvent insuffisants pour

construire une phylogénie robuste. Ainsi la méthode la plus utilisée est la technique de "bootstrapping", pour laquelle on génère plusieurs fois des estimations, à partir de l'échantillonnage de base, en utilisant la même méthode. Il est alors possible d'assigner à chaque branche un score représentant le nombre de fois où l'on retrouve cette même branche (Sanderson, 1989). Les branches faiblement supportées et conduisant à une histoire évolutive trop complexe au cours de la réconciliation doivent alors être mise en doute (Figure 3.10, D, E). Dans ces situations il est souvent plus juste d'opter pour l'histoire évolutive la plus parcimonieuse.

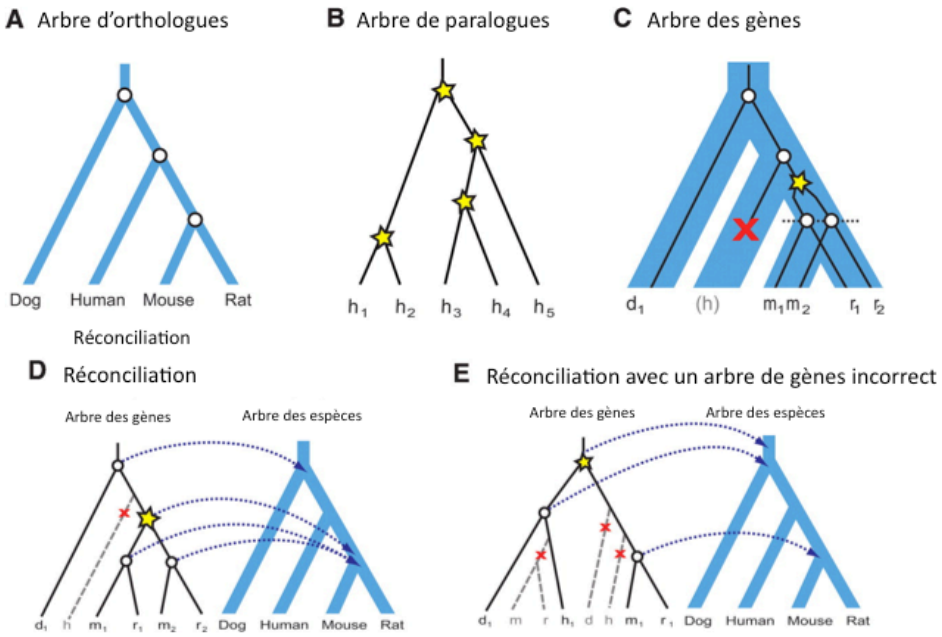


Figure 3.10 : Relation entre les arbres de gènes et les arbres d'espèces. (A) Arbres d'orthologues utilisés pour étudier l'évolution des espèces. Chaque nœud interne représente un événement de spéciation (cercle). (B) Arbres de paralogues utilisés pour étudier l'expansion de la famille de gènes au sein d'une même espèce. Chaque nœud interne représente un événement de duplication (étoile). (C) Arbre de gènes combinant les orthologues et les paralogues pour plusieurs espèces pour inférer les évènements de duplications (étoile) et de perte de gènes (croix), et la spéciation (cercle). Chaque gène est nommé par la première lettre de l'espèce correspondante. L'arbre de gène (lignes noires) peut être considérée comme évoluant à l'intérieur de l'arbre des espèces (zone bleue), impliquant des événements de spéciation au niveau des embranchements dans l'arbre des espèces (en pointillés). (D) Les duplications et les pertes de gènes sont déduits en réconciliant l'arbre de gène avec l'arbre des l'espèces, en faisant correspondre chaque nœud de l'arbre des gènes au nœud de l'ancêtre commun le plus proche de l'arbre des espèces (flèches). (E) Lorsque l'arbre de gène est incorrect, de nombreux événements parasites seront déduits. Dans cet exemple, une mauvaise position des rongeurs due à l'attraction des longues branches conduit à quatre événements artéfactuels : une duplication et au moins trois pertes de gènes. D'après (Rasmussen and Kellis, 2007).

III. Stratégie

1 La conservation des protéines

Afin de déterminer la conservation des protéines impliquées dans le métabolisme lipidique nous avons calculé les pourcentages d'identité et de similarité pour chaque paire d'orthologues pour chaque paire d'espèces et pour chaque paire de paralogues au sein d'une même espèce. Ces valeurs permettent de déterminer les gènes ayant une évolution rapide ou a contrario une forte conservation pour chaque espèce et dans le clade. Je présenterai principalement les données de similarité car elles rendent mieux compte de l'évolution entre deux protéines puisque la similarité tient compte des remplacements conservatifs.
Les pourcentages d'identité et de similarité ont ensuite été rapportés à la divergence moyenne d'un set de gènes communs à tous les génomes pour déterminer ce que nous appelons le ratio de similarité et le ratio d'identité.

$$\text{Ratio de similarité} = \frac{\%\text{ similarité pour l'alignement du gène Z pour les espèces X et Y}}{\%\text{ similarité pour l'alignement des 912 gènes pour les espèces X et Y}}$$

$$\text{Ratio d'identité} = \frac{\%\text{ identité pour l'alignement du gène Z pour les espèces X et Y}}{\%\text{ identité pour l'alignement des 912 gènes pour les espèces X et Y}}$$

Le set de gènes communs correspond aux 912 gènes utilisés pour réaliser la phylogénie des espèces séquencés du clade (voir encadré page 73) auxquels les gènes impliqués dans le métabolisme lipidique sont enlevés. Ces 912 gènes sont par définition des gènes bien conservés puisqu'ils sont présents dans les 10 génomes de levures disponibles sur Génolevures, dans le clade *Yarrowia* et dans OLHI et les orthologues partagent au moins 50% de similarité et s'alignent sur au moins 100 acides aminés.
Ces ratios permettent d'une part de corriger le biais dû à la divergence des espèces, afin que les valeurs obtenues pour toutes les paires d'espèces soient comparables malgré les différentes distances évolutives qui les séparent, et d'autre part de comparer l'évolution du couple de gènes à celle des génomes. Un ratio supérieur à 1 signifie que les protéines étudiées sont plus

fortement conservées que le reste du génome alors qu'un ratio inférieur à 1 montre que les protéines ont divergé plus rapidement.

2 Les pressions de sélection

Pour estimer les pressions de sélection agissant sur les protéines du métabolisme lipidiques, les dN/dS pour chaque paire d'orthologues et pour chaque paire de paralogues ont également été calculés. Les dN/dS sont calculés par la méthode de maximum-likelihood (ML) développée par Goldman et Yang (Goldman and Yang, 1994) en utilisant le package CodeML de PAML 4 (Yang, 2007). Ce modèle considère le processus de substitution entre les codons et non pas entre nucléotides ce qui diminue au maximum les biais et qui permet de prendre en compte les substitutions multiples en un même site.
Pour comparer les pressions de sélection s'exerçant sur les gènes du métabolisme lipidique avec celles s'appliquant sur les génomes complets nous calculons les dN/dS exclusivement pour chaque paire d'orthologues synténiques pour chaque couple d'espèces, en excluant les paires contenant un gène du métabolisme lipidique. Comme nous avons vu précédemment il existe beaucoup de familles multigéniques dans le clade Yarrowia donc nous préférons restreindre l'analyse aux gènes dont l'orthologie est confirmée par la synténie plutôt que de prendre le risque d'apparier des gènes ne dérivant pas du même gène ancestral. Si, pour un couple d'espèces données, la moyenne des dN/dS des orthologues synténiques est supérieure au dN/dS d'un gène étudié, ceci signifie que ce gène subit en moyenne plus de pression purifiante que le reste du génome.

Les premiers résultats de dN/dS obtenus n'étaient pas concluant voire absurdes. Nous avons alors étudié les distributions des dS, dN et dN/dS. Les courbes représentant la densité des distributions des dS montrent que certaines valeurs de dS sont aberrantes. Pour une part cela provient d'une limite de CodeML qui remplace les valeurs de dS par 99 pour éviter les divisions par 0 dans le cas où il n'y a pas de substitution synonyme. Il apparaît également qu'il faut enlever les couples de séquences dont le nombre de différences en nucléotides est égal à 0 ou 1 pour éviter tout calcul qui n'aurait pas de sens.
Mais de façon générale les valeurs aberrantes de dS obtenues sont principalement dues à la saturation des dS car s'il y a trop de substitutions synonymes possibles en un même site le signal est saturé. En partant du principe que pour un dS > 1 tous les sites silencieux ont étés substitués en moyenne plus d'une fois, on comprend qu'a partir de cette valeur la divergence calculée entre les séquences est très dépendante des hypothèses du modèle de substitution utilisé. Nous avons alors décidé de comparer les résultats obtenus par deux méthodes de calculs des dN/dS provenant du package PAML : Yn00 et CodeML. Nous observons alors que les dN obtenus sont

identiques alors que les dS sont différents. Les dS ne sont plus corrélés lorsque dS > 1. Nous choisissons donc d'exclure tous les couples pour lesquels il y a plus d'une substitution synonyme par site synonyme, donc lorsque dS > 1.

3 Reconstruction des scénarios évolutifs des familles de gènes

Dans un premier temps, pour chaque famille multigénique nous avons retracé la phylogénie afin de déterminer le nombre de gènes ancestraux avant la divergence du clade et le nombre d'évènements de duplication, de perte et de pseudogénisation. Pour retrouver ces évènements évolutifs nous avons utilisé une approche de parcimonie et nous nous sommes appuyés sur l'étude de la synténie. L'environnement de tous ces gènes dans les 6 génomes a été étudié afin d'observer la synténie dans le clade et de retrouver d'éventuels évènements de réarrangements. Si la synténie est conservée dans le clade ceci permet de confirmer qu'un gène est ancestral ou qu'il a divergé avant une duplication. Ceci permet aussi de donner des indices sur le type de duplication mis en jeu, les duplications récentes en tandem pourront par exemple être observées. Afin de retracer tous ces évènements évolutifs de manière homogène et de façon non ambigüe sur le sens acquisition/perte, nous n'avons considéré que les évènements ayant affecté le clade après la divergence de OLHI, qui sert donc de référence externe, et juste avant la divergence de YAAL (Figure 3.11). En effet, tous les évènements ayant eu lieu avant semblent très nombreux et il est très difficile de les retracer même avec une approche de parcimonie. La topologie des espèces utilisée comme référence pour les reconstructions des scénarios évolutifs est celle présentée dans l'encadré page 73 basée sur la concaténation de 912 protéines.

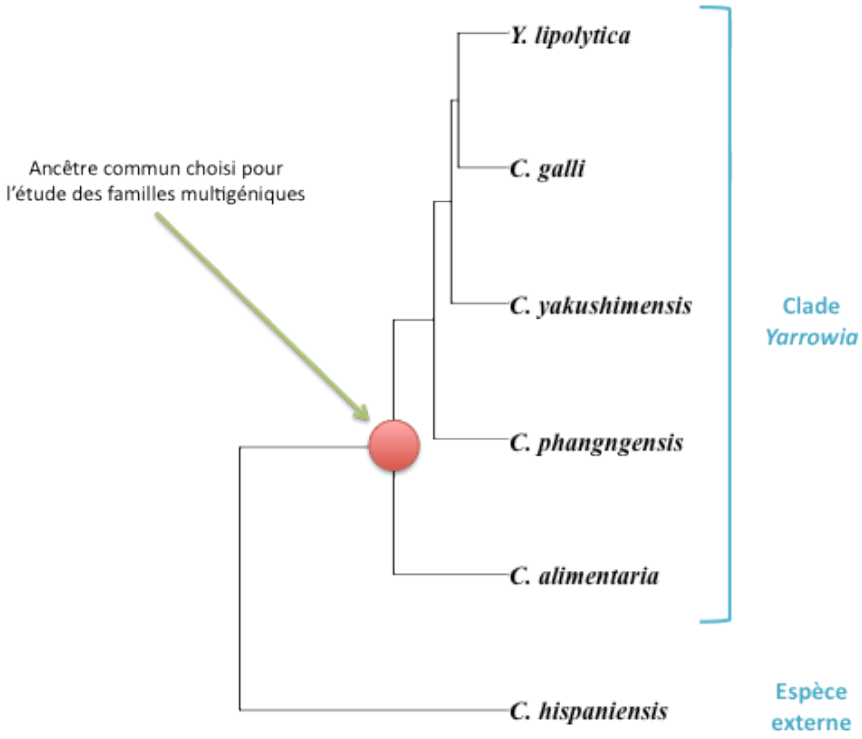


Figure 3.11 : Cladogramme du clade *Yarrowia*. Le rond rouge représente l'ancêtre commun choisi pour la reconstruction des évènements évolutifs au sein des familles multigéniques.

Dans un second temps nous avons trié les familles selon un indice, que nous appelons Dyn, correspondant au ratio du nombre d'évènements sur le nombre de gènes ancestraux.

$$\text{Dyn} = \frac{\text{Nombre d'évènements (duplications + pertes + pseudogénisation)}}{\text{Nombre de gènes ancestraux}}$$

Une famille dont le ratio est égal à 0, c'est à dire qu'aucun évènement n'a été observé, est appelée famille gelée. Une famille avec un ratio supérieur à zéro est considérée comme dynamique car elle aurait subi au moins un réarrangement depuis la divergence de OLHI. Parmi les familles dynamiques nous distinguons des familles peu dynamiques, avec un Dyn compris entre 0 et 1, et les familles très dynamiques avec un Dyn supérieur ou égal à un, c'est à dire des familles qui auraient subi au moins un réarrangement par gène ancestral.

Après avoir reconstruit ces scénarios évolutifs, comme décrit précédemment et schématisé dans la figure 3.12, nous avons regroupé ensemble tous les descendants d'un même ancêtre commun afin de constituer des groupes d'orthologues. Nous avons ensuite étudié de manière fine l'évolution des protéines au sein des familles en étudiant leur conservation et leurs pressions de sélection en utilisant les stratégies expliquées dans les paragraphes 2.1 et 2.2.

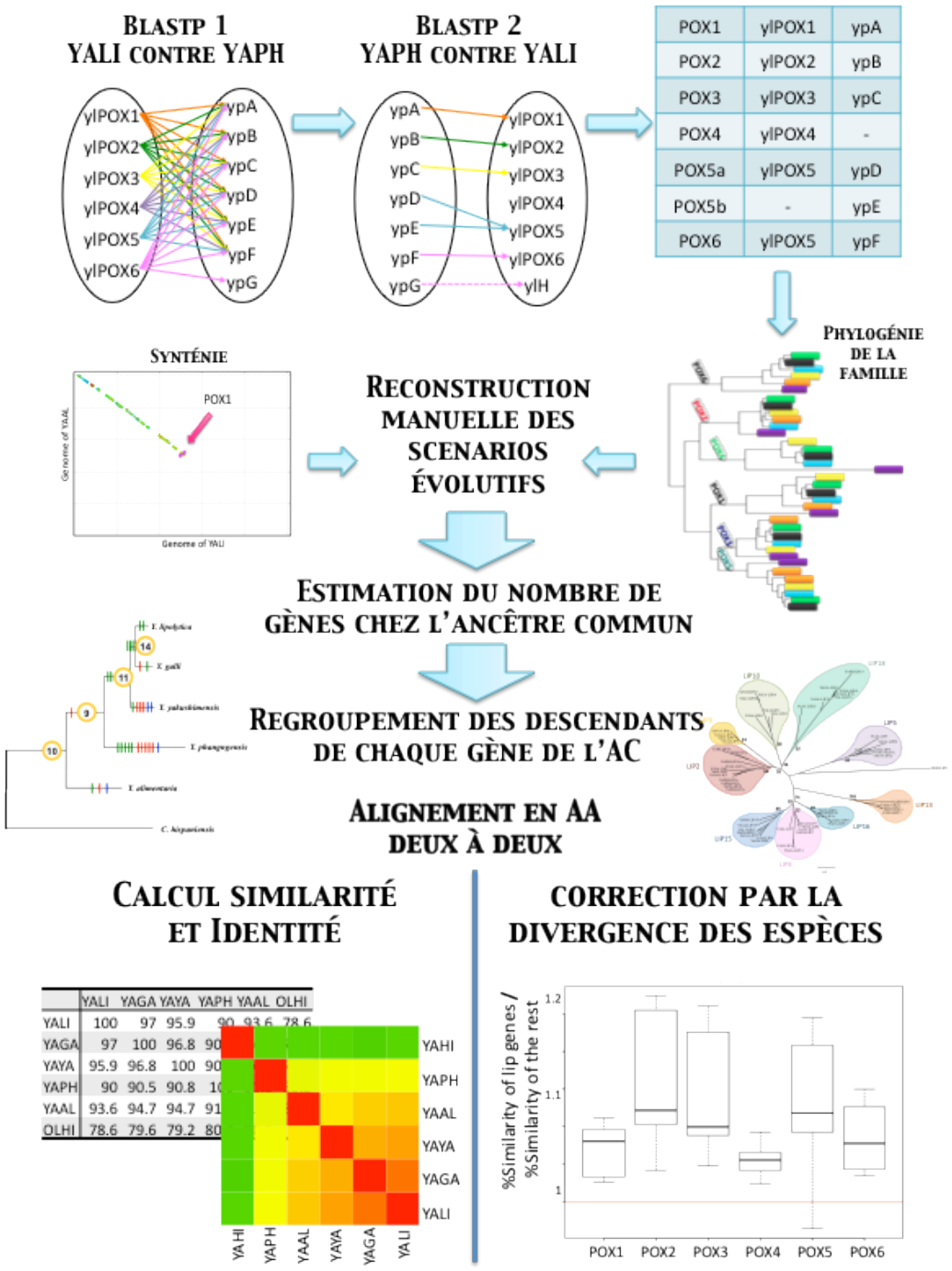


Figure 3.12 : Stratégie utilisée pour la reconstruction des scénarios évolutifs des familles de gènes. Ici est donné l'exemple de la famille POX.

4 Stratégie des scripts

4.1 Script 1 : dN/dS

Ce script calcule les dN/dS des paires d'orthologues synténiques pour le clade et les dN/dS des paires de gènes du métabolisme lipidique dont nous savons qu'ils proviennent d'un même ancêtre commun après reconstruction des scénarios évolutifs (Figure 3.13).

Pour déterminer les paires de gènes d'orthologues synténiques nous utilisons l'outil SynChro (Drillon et al., 2014). SynChro détermine les reciprocal best BLAST hit (RBH) avec Opscan (http://wwwabi.snv.jussieu.fr/public/opscan/) et ne conserve que les RBH en synténie. Pour chaque paire de gènes orthologues et pour chaque paire d'espèce, les deux séquences protéiques sont alignées avec mafft (Katoh et al., 2002). Les pourcentages d'identité et de similarité sont calculés sur la longueur de l'alignement avec aligncopypair de la suite EMBOSS (Rice et al., 2000) et le script détermine le nombre de différences en acides aminés, le nombre de gaps et la longueur de l'alignement. Ensuite l'alignement en nucléotides est déduit de l'alignement en acides aminés en utilisant tranalign de la suite EMBOSS (Rice et al., 2000) et, comme pour l'alignement en acides aminés, les pourcentages d'identité et de similarité ainsi que le nombre de différences, le nombre de gaps et la longueur de l'alignement sont calculés. Enfin, les dN, les dS et les dN/dS sont calculés à partir de cet alignement nucléotidique en utilisant le package CodeML de PAML4 (Yang, 2007). Les paires pour lesquelles au moins un des gènes est inférieur à 100 acides aminés sont supprimées car sur des gènes de trop petite taille les différences sont surévaluées. De plus, nous avons décidé d'autoriser au maximum une substitution observable S ou N par codon, c'est à dire que le nombre de différences en nucléotides doit être inférieur au nombre de codons, pour être sûr de comparer des gènes suffisamment proches. Et enfin, comme expliqué dans le paragraphe 2.2 nous avons décidé d'enlever les couples pour lesquels le dS>1. Les résultats de dN/dS présentés dans ce chapitre sont à alors à relativiser car après avoir trié selon ces critères seulement 14315 paires d'orthologues synténiques sur 51802 ont pu être considérées dans les analyses.

Le script concatène alors les résultats pour chaque paire de gènes dans un tableau. Le script teste pour chaque paire d'espèces si les valeurs de dN/dS des gènes du métabolisme lipidique sont significativement différentes des valeurs obtenues pour les autres gènes. Nous utilisons un test de Student qui compare les moyennes car les données ne sont assez nombreuses pour appliquer un test du Chi2. L'hypothèse nulle H_0 est l'égalité des moyennes des deux échantillons.

Le script génère également 6 types de graphiques :

- Pour chaque couple d'espèces les courbes représentant la densité des distributions des dN/dS en rouge pour les gènes du métabolisme lipidique et en noir pour les autres gènes.
- Pour chaque couple d'espèces un premier boxplot des dN/dS en log des gènes du métabolisme lipidique et un deuxième pour les autres gènes.

Ces deux types de représentations côte à côte, distributions et boxplots, permettent d'observer la répartition des données mais aussi de comparer les distributions des deux sets de gènes et d'avoir un *a priori* sur la différence des dN/dS entre les gènes du métabolisme lipidique et les autres.

- Pour chaque couple d'espèces un boxplot par famille de gènes du métabolisme lipidique contenant des valeurs de dN/dS. Les boxplot sont classés par ordre croissant du nombre de gènes dans la famille présents parmi les couples d'orthologues synténiques. Ceci permet de visualiser facilement si certaines familles ont une évolution différente des autres.
- Pour chaque couple d'espèces, les boxplots du dN/dS en fonction du nombre de gènes dans la famille présents parmi les couples d'orthologues synténiques. Par exemple toutes les valeurs de dN/dS correspondant aux gènes appartenant à une famille de trois membres sont réunies dans le même boxplot. Les boxplots sont alors classés par ordre croissant ce qui permet de repérer rapidement si le dN/dS varie en fonction du nombre de gènes dans les familles.
- Pour chaque couple d'espèces, les dN/dS pour chaque gène impliqué dans le métabolisme lipidique.
- Pour toutes les espèces, les boxplot des médianes des dN/dS pour chaque gène impliqué dans le métabolisme métabolique. Dans cette représentation ne sont conservés que les gènes présents dans au moins trois paires d'espèces.

Dans ces deux derniers types de graphiques les données sont triées par ordre croissant afin de visualiser rapidement les gènes avec la plus forte sélection positive ou la plus forte sélection purifiante.

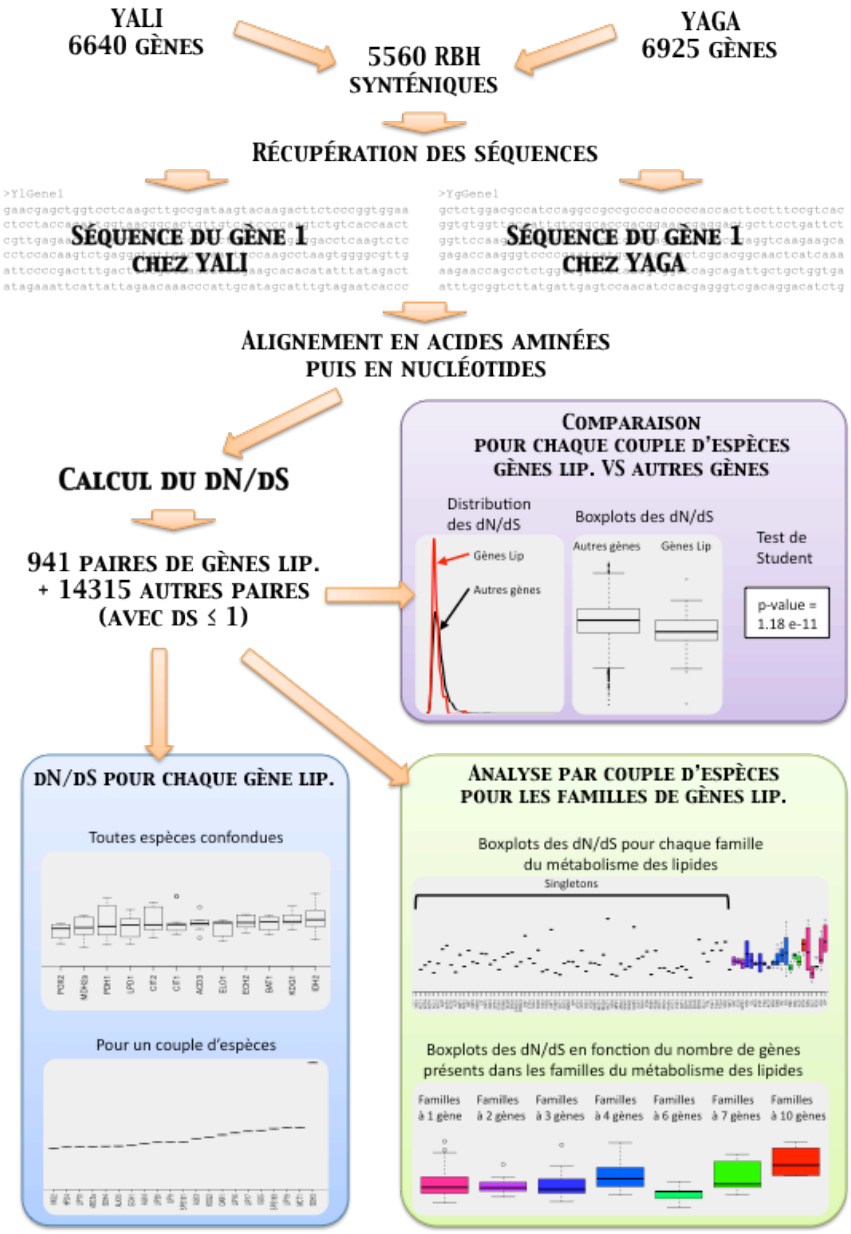


Figure 3.13 : Stratégie du script permettant d'évaluer les pressions de sélection sur les gènes du métabolisme lipidique et de les comparer aux autres gènes des génomes. Dans ce schéma l'exemple est donné pour le couple YALI-YAGA.

4.2 Script 2 : Evolution des familles du métabolisme lipidique

Pour étudier l'évolution des familles du métabolisme lipidique dans le clade *Yarrowia* j'ai développé un script qui prend en entrée les tableaux de gènes impliqués dans le métabolisme lipidique construits pour chaque génome du clade pour l'étude précédente (voir Chapitre 2), ainsi qu'une liste contenant tous les noms des gènes d'intérêts classés en famille fonctionnelles et les fichiers multifasta en acides aminés avec les séquences de ces gènes.

Le script réalise alors les alignements multiples en acides aminés à partir de mafft (Katoh et al., 2002). Comme expliqué précédemment, deux types d'alignements sont calculés : un pour les

paralogues et un pour les orthologues dans le clade. Les pourcentages d'identité, de similarité et les dN/dS sont calculés à partir des paralogues d'une part et des orthologues d'autres part, en utilisant aligncopypair et tranalign de la suite EMBOSS (Rice et al., 2000), et le package CodeML de PAML4 (Yang, 2007). Les pourcentages globaux d'identité et de similarité sont également calculés pour toutes les paires d'espèces à partir des alignements des 912 gènes de chaque génome codant des protéines, présents en en copie unique dans chaque génome (398959 résidus) (voir encadré page 73) afin de déterminer les ratios de similarité et d'identité.
Le script fournit toutes ces données sous forme de tableaux. Ensuite il organise ces résultats en plusieurs matrices symétriques contenant les pourcentages d'identité et de similarité, les ratios correspondant et les dN/dS pour chaque paire d'espèces pour un gène ou pour chaque paire de gènes d'une famille pour une espèce. Les matrices servent alors de support pour dessiner des heatmaps offrant une visualisation rapide des différences d'évolution qu'il y a dans le clade pour le métabolisme lipidique.
Le script génère également 4 types de graphiques :
- Pour chaque gène un boxplot contenant les valeurs des ratios de toutes les paires d'espèces est dessiné. Ces valeurs proviennent des alignements de gènes orthologues dans le clade. Les boxplots sont alors classés par ordre croissant ce qui permet de repérer rapidement les gènes qui évoluent très peu ou très fortement.
- Pour chaque famille multigénique, on représente un boxplot par gène contenant les valeurs des ratios ou des pourcentages de toutes les paires d'espèces. Ces valeurs proviennent des alignements de gènes orthologues dans le clade. Cette représentation permet de visualiser facilement si un gène dans une famille évolue différemment.
- Pour chaque famille multigénique, le script fait un boxplot par espèce contenant les valeurs des ratios ou des pourcentages de toutes les paires de gènes de cette famille. Ces valeurs proviennent des alignements des paralogues au sein de chaque espèce. Ceci permet de repérer rapidement les espèces pour lesquelles les gènes d'une même famille sont très divergents ou au contraire s'ils restent très semblables.

4.3 Script 3 : La synténie

Pour comprendre les mécanismes impliqués dans les évènements de duplication, pertes ou pseudogénisation, nous avons décidé de chercher des arguments dans la conservation de la synténie. J'ai alors écrit un script qui permet tout d'abord de récupérer les séquences des gènes voulus avec un nombre défini de base en aval et en amont (Figure 3.14). Dans notre cas le script est utilisé pour récupérer toutes les séquences des gènes du métabolisme lipidique en famille multigénique avec 50000 bases de chaque coté. Puis, pour chaque espèce toutes ces séquences sont alignées contre les orthologues de YALI afin de dessiner un dotplot par paire en faisant

appel à Promer de MUMmer v3.23 pour calculer les coordonnées des régions en synténie et à mummerplot de MUMmer v3.23 pour tracer les dotplots (Kurtz et al., 2004). Le script détermine ensuite pour chaque paire s'il y a conservation de la synténie entre les régions contenant les gènes orthologues. Pour cela il calcule la somme des longueurs des matchs, déterminées par Promer. J'ai décidé de façon arbitraire qu'il n'y a pas de synténie si cette somme est inférieure à 6 fois la longueur du gène car l'alignement pour le dotplot est fait dans les 3 cadres de lecture soit en sens soit en antisens, donc on considère qu'il n'y a pas de synténie s'il n'y a pas au moins une seconde région de la taille du gène de référence qui s'aligne.

S'il n'y a pas de synténie, le script va générer des dotplots de la région étudiée contre chacun des chromosomes de l'espèce de référence, afin de repérer quelle(s) région(s) est(sont) en synténie avec le gène. Pour tous les dotplots, le script modifie les fichiers de coordonnées de Promer afin de marquer la position du gène sur le graphique.

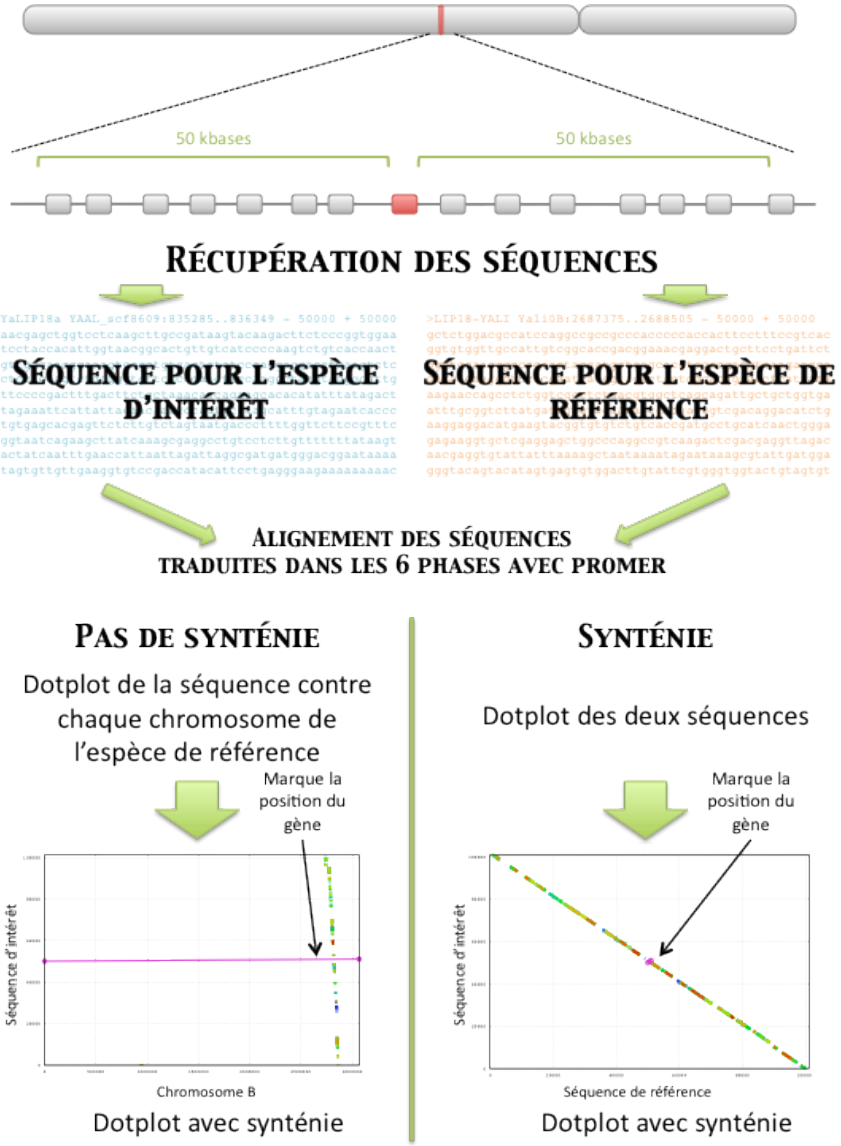


Figure 3.14 : Stratégie du script permettant d'étudier la synténie d'une liste de gènes avec les orthologues d'une espèce de référence.

IV. RÉSULTATS

1 VUE D'ENSEMBLE DE L'ÉVOLUTION DE CE MÉTABOLISME

1.1 Existent-ils des voies du métabolisme lipidique qui évoluent différemment ?

Le but est de rechercher si certains gènes ou certaines voies particulières du métabolisme lipidique sont de façon globale plus contraintes que le reste du métabolisme lipidique dans le clade. Ainsi, pour toutes les espèces confondues, nous calculons la moyenne des dN/dS des gènes descendant d'un même gène ancestral du métabolisme lipidique et nous enlevons les gènes représentés par moins de 3 paires d'espèces car la moyenne ne serait pas significative. Nous obtenons donc un set de 125 gènes dont 65 appartiennent à des familles. De façon surprenante les dN/dS sont très faibles. Ceci suggère que soit tous ces gènes sont soumis à une pression de sélection purifiante, soit les méthodes de calcul du dN/dS ne sont pas adaptées à ce clade de levures.

Parmi les 10 gènes dont la moyenne des ratios dN/dS est la plus faible, allant de 0,015 à 0,030, deux gènes codent pour des protéines impliquées dans la synthèse d'ester de stérol, GCY16 et GPD1, deux gènes codent la réaction réversible qui va du pyruvate à l'oxaloacétate, PDC1 et PYC1. Les autres gènes sont CRC1 impliqué dans la navette carnitine, ACL1 dans la navette malate/citrate, FAS1 dans l'élongation des acides gras, LSC1 dans le cycle du TCA, ACS2 qui permet l'activation des acides gras et ERG10 pour la synthèse de l'acétyl-CoA. Le gène ayant le plus faible dN/dS, dN/dS = 0,016, est CRC1 qui code un transporteur de carnitine.
Parmi ces dix gènes, seuls CRC1 et GCY16 font partie d'une famille multigénique. GCY16 fait partie d'une famille à 10 membres chez YALI et CRC1 appartient à la famille YMC qui compte 4 membres chez YALI.
Le gène ayant le dN/dS le plus élevé, dN/dS = 0,24, est SDH3 qui code une des quatre sous-unités constituant le tétramère de la succinate dehydrogenase requise dans le cycle du TCA. Parmi les dix gènes ayant les dN/dS les plus élevé, trois participent au cycle du TCA : SDH3, SDH4 et KGD2, deux gènes sont des ALK, requises pour l'assimilation des alcanes, il y a le gène

ISO5 proposé comme enzyme auxiliaire à la ß-oxydation, une lipase LIP13, FAT3 transporteur d'acides gras, DAK1 impliqué dans la synthèse d'ester de stérols et MGH1 qui permet la synthèse d'acétyl-CoA par la dégradation de la leucine et potentiellement requis dans la ß-oxydation mitochondriale. Huit de ces 10 gènes appartiennent à des familles multigéniques d'au moins trois gènes chez YALI. Ces familles comptent les LIP et les ALK qui sont les plus grandes familles du métabolisme lipidique chez YALI. Les moyennes des dN/dS sont effectivement corrélées au nombre de gènes dans la famille (test de corrélation : Pearson : p-value = 0,023 ; corrélation = 0,2).

Lorsque l'on observe les 10 gènes avec le ratio de similarité le plus élevé, nous retrouvons, comme pour les dN/dS les plus élevés ACL1 et PYC1. Il y a également GUT1 impliqué dans la synthèse d'ester de stérols, PDB1 dans la synthèse d'acétyl-CoA, la pyruvate kinase PYK1 et l'alcool déshydrogénase ADH1. Ce dernier gène n'est pas dans le jeu de données utilisé pour l'analyse des dN/dS car un seul couple a un dS<1. Les quatre gènes restants sont tous requis dans le cycle du TCA : ACO1, IDH1, CIT2 et PDH1. Dans ces 10 gènes, 3 appartiennent à des familles de 5 membres pour les ADH, 3 pour les ACO et 2 pour les CIT.
Parmi les dix gènes ayant le ratio de similarité le plus faible, six ne font pas partie des données de l'analyse du dN/dS car ils sont trop divergents entre les espèces du clade. Ce sont ALK11, les LIP4, LIP5 et LIP8, le singleton AGX1 et COU10 qui appartient à la famille des 10 coumarate-CoA ligases. Nous retrouvons alors comme pour les plus forts dN/dS une majorité de gènes appartenant aux familles nombreuses notamment LIP et ALK. Les quatre autres gènes sont la LIP2, les deux facteurs de transcription YAS2 et YAS3 et la fatty acyltransferase SLC1 requise dans la synthèse d'esters de stérol.

De manière générale, nous pouvons observer que le ratio de similarité est négativement corrélé au dN/dS (test de corrélation : Pearson : p-value = 7.466e-08 ; corrélation = -0,46). Ainsi, la plupart des gènes ayant les dN/dS les plus faibles ont également un fort ratio de similarité (Figure 3.15). Le nombre de gènes dans les familles est corrélé au dN/dS (test de corrélation : Pearson : p-value = 0,02) et au ratio de similarité (test de corrélation : Pearson : p-value = 0,008).

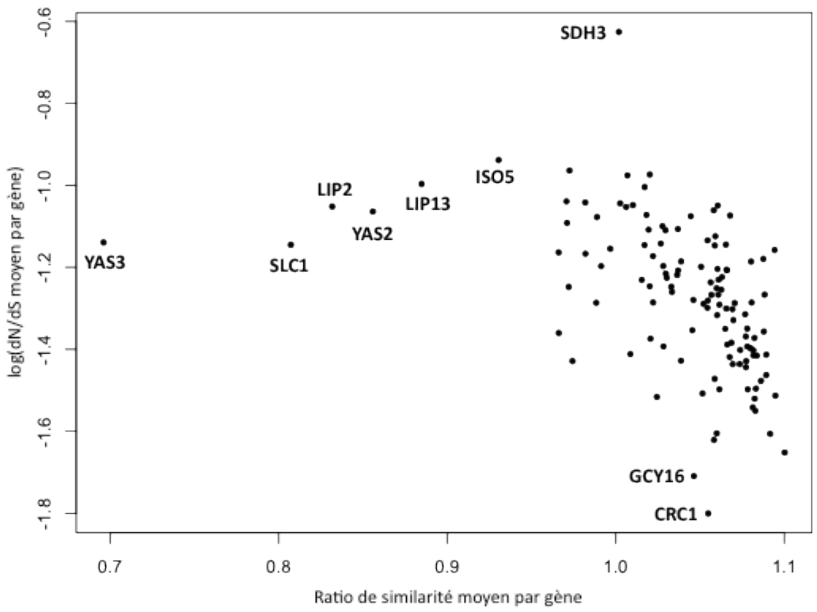


Figure 3.15 : La moyenne des ratios de similarité en fonction de la moyenne du dN/dS pour chaque gène du métabolisme lipidique pour les espèces du clade *Yarrowia*. Le dN/dS est représenté en log. Les noms des gènes qui ont une valeur de dN/dS ou un ratio de similarité extrême sont notés à coté des points correspondants.

1.2 Gènes du métabolisme lipidique vs. génomes complets

Pour déterminer les pressions de sélection qui s'exercent sur les génomes du clade, les dN/dS sont calculés sur 14315 paires d'orthologues synténiques obtenues pour les dix couples d'espèces, n'incluant pas OLHI.

Parmi les 14315 paires de gènes analysées, une paire seulement a un dN/dS supérieur à 1, dN/dS = 1,12. Cette paire est composée du gène YALI0C15741g et son homologue chez YAGA YAGA0B11122g. Lorsque l'on réalise le blast de YALI0C15741g contre la base non redondante (nr) du NCBI les quatre premiers hits sont des protéines de YALI et les autres ne partagent que très peu de similarité (avec des e-value < 0,9). Le gène YALI0C15741g compte 10 paralogues dont six pseudogènes annotés sans similarité (YALI0B06138g, YALI0A14927g, YALI0C15785g, YALI0F15675g et les pseudogènes : YALI0D16016g, YALI0A10461g, YALI0A01507g, YALI0A09361, YALI0A00297, YALI0A01584g) et YAGA0B11122g à deux pseudogènes paralogues (YAGA0E26720g et YAGA0A12288g). Il n'existe aucune séquence homologue à ces gènes dans le reste du clade suggérant que cette famille est apparue après la divergence de YAYA et qu'elle est en expansion chez YALI et YAGA.

Les gènes de cette famille sont très peu exprimés sur les milieux glucose, acide oléique et tributyrine chez la souche W29 de YALI. Par contre, nous avons pu observer chez la souche E150 de YALI, grâce à d'autres données de transcriptomique présentes dans le laboratoire, que les gènes YALI0B06138g, YALI0C15741g et plus particulièrement YALI0D16016g sont fortement exprimés sur alcane et très fortement en milieu H_2O_2, c'est à dire en condition de stress oxydatif (Tableau 3.1).

Le second dN/dS le plus élevé, dN/dS = 0,65, est obtenu pour le couple YALI0C21318g-YAYA0S4-03532g. Ces gènes n'ont aucune similarité avec les gènes de la base non redondante du NCBI mais possèdent le domaine DUF2420 lui même sans fonction connu mais annoté comme conservé chez les champignons. Ces gènes ont un orthologue dans chaque génome du clade (YAGA0B14862g, YAPH0S6-07580g et YAAL0S01-00848g) mais pas chez OLHI suggérant son apparition après la divergence de OLHI.

Pour l'ensemble des paires, 75% ont un dN/dS inférieur à 0,09. Pour les 941 paires de gènes impliqués dans le métabolisme des lipides, 75% des ratios sont inférieurs à 0,07. En moyenne les dN/dS des gènes du métabolisme sont plus faibles que ceux des autres gènes du génome, $dN/dS_{moyenLip}$ = 0.059 et $dN/dS_{moyenGénome}$ = 0,072 (Figure 3.16). De même, les ratios de similarités, obtenus pour 1997 couples de gène lipidiques n'incluant pas OLHI, sont supérieurs à 1 pour plus de 70% des ces couples. Ceci signifie qu'ils sont majoritairement plus conservés que les autres gènes des génomes.

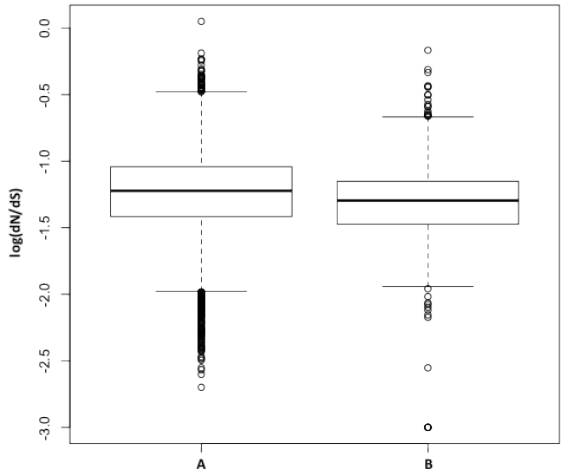


Figure 3.16 : Boxplot des log(dN/dS) de toutes les paires de gènes pour tous les couples d'espèces du clade *Yarrowia* (OLHI n'est pas inclus) A) dN/dS des orthologues synténiques n'appartenant au métabolisme lipidique. B) dN/dS des gènes du métabolisme lipidique.

Pour chaque paire d'espèces les moyennes des dN/dS obtenues pour les gènes impliqués dans ce métabolisme sont toujours plus faibles que celles obtenues pour le reste des gènes (Figure 3.17). Ceci est d'autant plus vrai pour les deux couples d'espèces les plus proches, c'est à dire YALI-YAGA et YALI-YAYA. Pour ces deux couples les p-values obtenues avec un test de Student sur les log des dN/dS sont de 1.556e-14 et 2.361e-05 respectivement, ce qui permet de conclure que les moyennes des dN/dS des deux échantillons sont significativement différentes. Pour les autres couples le test n'est pas significatif mais la même tendance est observée. Ceci suggère que les gènes du métabolisme étudié subissent une sélection purifiante plus forte que le reste du génome.

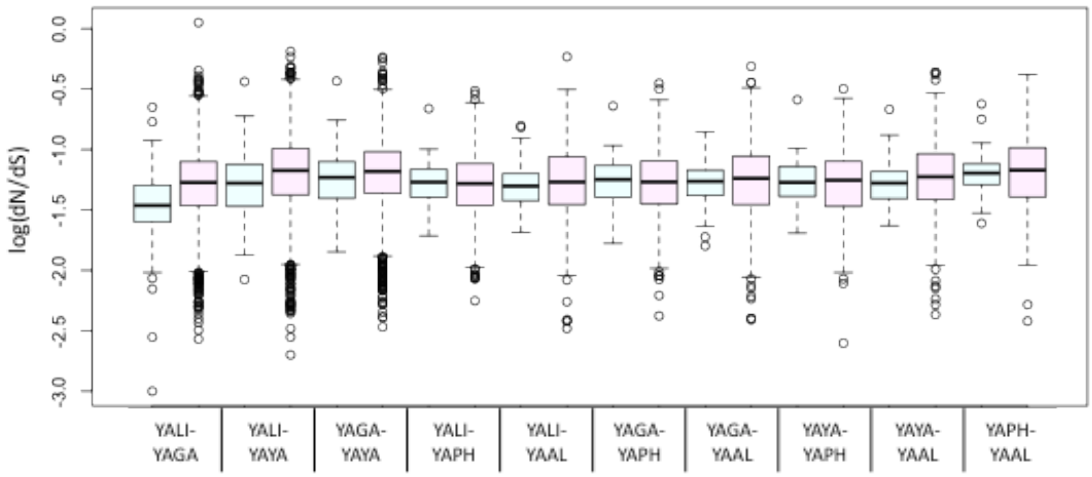


Figure 3.17 : Pour chaque couple d'espèces du clade *Yarrowia* : boxplot des log(dN/dS) de toutes les paires de gènes du métabolisme lipidique en bleu et des orthologues synténiques n'appartenant pas à ce métabolisme en rose.

Tableau 3.1 : Expression relative des gènes de la famille contenant le couple de gènes avec le dN/dS le plus élevé. Les données proviennent des transcriptomes déposés à l'ENA-EMBL sous le numéro de projet PRJEB6632 (voir article page 163), pour les milieux glucose, acide oléique et tributyrine, et de transcriptomes de l'équipe, non présentés, pour les milieux alcane et H_2O_2. Les gènes en gris correspondent aux pseudogènes.

	Glucose	**Acide Oléique**	**Tributyrine**	**Alcane**	**H_2O_2**
YALI	**Souche W29**			**Souche E150**	
YALI0A00297g	-/+	-/+	-/+	-	-
YALI0A01507g	-	-	-	-	-
YALI0A01584g	-	-	-	-	-
YALI0A09361g	-	-	-	-/+	-/+
YALI0A10461g	-	-	-/+	-	-
YALI0A14927g	-	-	-	-	-
YALI0B06138g	-	-	-/+	++	+++
YALI0C15741g	-	-	-/+	+	+++
YALI0C15785g	-	-	-/+	-	-/+
YALI0D16016g	-	-	+	+++	+++++
YALI0F15675g	-	-	-/+	-	-
YAGA	**Glucose**	**Acide Oléique**	**Tributyrine**	**Alcane**	**H_2O_2**
YAGA0A12288g	-/+	-	-	NA	NA
YAGA0B11122g	-/+	-	-	NA	NA
YAGA0E26720g	-/+	-/+	-/+	NA	NA

2 Evolution des familles de gènes

Chez YALI, la capacité oléagineuse est facilitée par l'expansion de familles de protéines ayant eu lieu après la divergence du clade *Yarrowia*. Nous avons vu par exemple dans le chapitre 2 que la famille des lipases contient 16 membres chez YALI contre seulement un chez SACE. Parmi les 204 gènes de YALI impliqués dans le métabolisme lipidique, 136 sont répartis dans 37 familles multigéniques, ce qui représente plus de 66% de ces 204 gènes et cette répartition des gènes en familles est très semblable chez les autres espèces du clade. Ces familles de gènes sont composées de groupes de gènes très similaires, reflétant l'hypothèse que tous proviendraient d'un même ancêtre commun. Les familles de gènes sont en effet la conséquence directe du fait que la plus grande partie des nouveaux gènes proviennent de duplication de tout ou partie du génome comprenant des gènes entiers. Les familles de gènes fournissent alors des informations sur la façon dont de nouveaux gènes apparaissent et se diversifient. De même, leur organisation peut fournir des indices sur la façon dont l'expansion des gènes dans certaines familles est régulée. Enfin, la présence de multiples copies de gènes apparentées peut permettre de mettre en évidence des forces évolutives non visibles par l'étude des gènes individuels.
Nous avons donc décidé d'étudier les familles de gènes impliquées dans le métabolisme lipidique au sein du clade *Yarrowia* afin de mettre en évidence les mécanismes évolutifs mis en jeu dans ce clade pour ce métabolisme.

2.1 Des familles dynamiques en expansion dans le clade

Nous avons vu dans le chapitre précédent que parmi les 204 gènes de *Y. lipolytica* impliqués dans le métabolisme lipidique, 136 sont répartis en 37 familles de gènes. Chez YALI, ces 37 familles sont composées de deux à 16 gènes et comptent en moyenne 3,68 gènes par famille (médiane = 2 gènes).
A partir des 37 phylogénies de ces familles multigéniques et de la synténie nous avons reconstitué manuellement les évènements probables de duplication, de perte de gène, de pseudogénisation et de perte de synténie, et nous avons déterminé le nombre de gènes présents chez l'ancêtre commun. Nous avons ainsi pu classer les familles en fonction de l'indice Dyn (Tableaux 3.2 et 3.3). Nous recensons ainsi 21 familles gelées (Dyn=0) et 16 familles dynamiques dont 7 familles peu dynamiques (0<Dyn<1) avec un Dyn moyen de 0,7 et 9 familles très dynamiques (Dyn≥1) avec un Dyn moyen de 1,4. Le Dyn maximal observé est celui de la famille des LIP, Dyn = 3,3 et le minimal est celui de la famille des POX, Dyn = 0,5.

Tableau 3.2 : Les familles gelées au sein du clade *Yarrowia* : bilan du nombre de gènes chez OLHI et chez l'ancêtre commun.

Nombre de gènes chez OLHI	Nombre de gènes ancestraux	Nom de la famille
2	2	ACP, ALD, ANT, BAT, CIT, ELO, ICL, IDH, MDH, MLS, PDA, POT, PXA, TGL
3	3	ACO, BKT, ECH, FAT
1	3	DAK
4	4	ACD, YMC

Tableau 3.3 : Les familles dynamiques au sein du clade *Yarrowia* : bilan du nombre de gènes chez OLHI et chez l'ancêtre commun, le nombre de perte de synténie, le nombre de duplications de gène, de pertes, de pseudogénisations et le ratio Dyn.

	Famille	Gènes OLHI	Gènes ancestraux	Perte de synténie	Dupli-cations	Pertes	Pseudo-génisation	Dyn
Peu dynamiques	POX	6	6	0	1	1	1	0,5
	ISO	2	5	1	0	2	1	0,6
	SPS	1	3	1	0	2	0	0,6
	COU	4	10	4	0	7	0	0,7
	ALK	7	12	4	2	5	2	0,75
	ABC	4	6	2	1	3	1	0,8
	ADH	4	6	3	0	5	0	0,8
							Moyenne :	0,7
Très Dynamiques	CNT2	1	1	0	1	0	0	1
	CNT4	1	1	0	1	0	0	1
	EST	1	4	0	1	3	0	1
	LSC2	1	1	0	1	0	0	1
	PDC1	1	1	0	1	0	0	1
	HFD	2	4	1	1	2	1	1
	SLC1	1	1	0	1	0	1	2
	GCY	5	5	2	7	2	1	2
	LIP	3	10	3	15	11	4	3
							Moyenne :	1,4
Total		44	76	21	33	43	12	
Moyenne		2,8	4,8	1,3	2,1	2,7	0,8	1,1

Les 21 familles gelées ont en moyenne 2,4 gènes ancestraux (médiane = 2) avec au minimum 2 gènes et au maximum 4. Les familles dynamiques ont en moyenne 4,8 gènes ancestraux (médiane = 4,5) avec au minimum un seul gène ancestral pour les SPS et au maximum 12 gènes pour les ALK. La médiane du nombre de duplications est de 1 et celle du nombre de pertes est de 2. Mais les duplications touchent en moyenne 1,5 espèces car sur les 33 duplications, 4 ont eu lieu avant la divergence de YALI-YAGA, 4 avant celle des *sensu stricto* et 1 avant celle de YAPH.

Par contre les 43 pertes sont toutes espèce spécifique, sauf dans trois cas où elles ont eu lieu avant la divergence de YAPH, ce qui fait une moyenne de 1,3 espèces touchées par une duplication. Le nombre de gènes en familles impliqués dans le métabolisme lipidique tend à augmenter suggérant une amplification de ces familles au sein du clade.

Pour plus de la moitié des familles dynamiques (9 cas sur 16), le nombre de gènes chez OLHI est inférieur à celui de l'ancêtre commun. Dans la plupart des cas ceci s'explique par des duplications qui ont eu lieu juste après la divergence de OLHI. Par exemple, le gène ancestral ISO2 s'est dupliqué juste après la divergence de OLHI pour donner le gène ISO1 dans le clade (Figure 3.18).

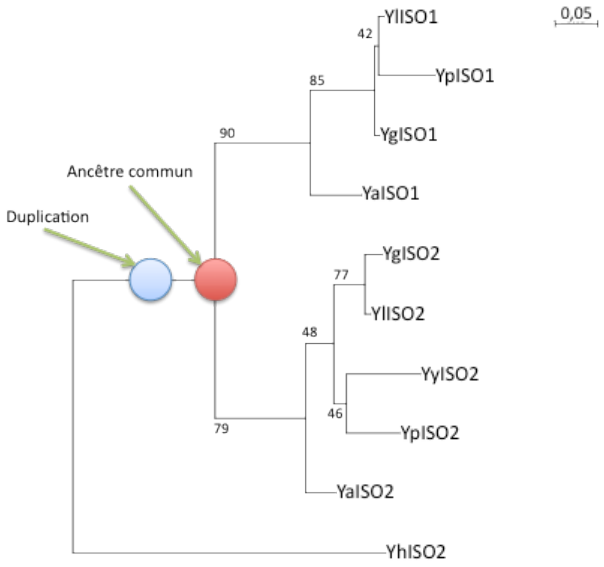


Figure 3.18 : Arbre phylogénique des gènes ISO1 et ISO2 pour le clade *Yarrowia* et OLHI provenant d'un alignement en acides aminés, 270 résidus. Le rond bleu représente la duplication avant l'ancêtre commun au clade. Le rond rouge représente l'ancêtre commun choisi pour la reconstruction des évènements évolutifs au sein des familles multigéniques. Le gène ISO1 est perdu chez YAYA.

Au cours de l'évolution des familles dynamiques il y a eu de nombreux évènements de réarrangements ayant conduit à une de perte de synténie des gènes dans le clade. Nous avons comptabilisé ces évènements car ils peuvent être le témoignage d'un déplacement du gène ou d'une duplication suivie d'une perte d'une des copies (tableau 3.3). Les pertes de synténie peuvent être espèces spécifiques ou avoir touché un ancêtre commun et dans ce cas elle est observable chez plusieurs espèces. Par exemple les gènes YaCOU8 et OhCOU8 sont en synténie entre eux mais pas avec les autres gènes COU8 du clade qui eux même sont en synténie (Figure 3.19). Ceci montre que le réarrangement ou la duplication suivie de la perte d'une copie a eu lieu après la divergence de YAAL.

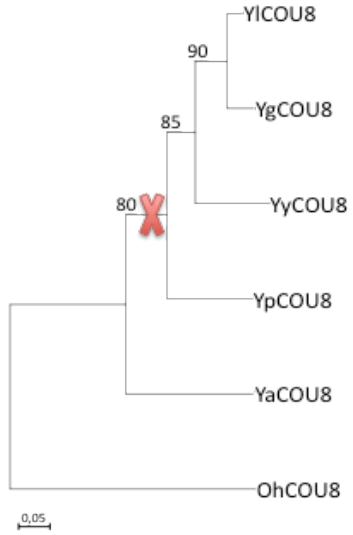


Figure 3.19 : Arbre phylogénétique des protéines COU8 appartenant à la famille très dynamique COU, 610 résidus. La croix rouge avant la divergence de YAAL marque la perte de synténie dans le clade.

Dans la famille ALK nous observons une perte de synténie du gène ALK11 après la divergence de YAPH et une après la divergence de YAAL pour le gène ALK7. La synténie du gène HFD3 a également été perdue après la divergence de YAAL. Tous les autres réarrangements observés sont espèce spécifiques.

2.2 Vers une contraction des familles ?

OLHI compte 44 gènes dans les familles dynamiques soit 36 de moins que YAAL, l'espèce la plus proche, et 41 de moins que YALI, l'espèce la plus éloignée. Par contre l'ancêtre commun que nous avons choisi pour l'analyse a seulement 10 gènes de moins que YALI. Ceci montre que l'expansion de ces familles de gènes a principalement eu lieu après la divergence de OLHI et avant l'apparition de l'ancêtre commun choisi.

Chez YAPH le nombre de gènes en familles dynamiques n'est que de 58 contre 85 pour YALI et 80 pour YAAL. C'est la seule espèce qui a moins de gènes en famille dynamique que l'ancêtre commun. Nous savons que YAPH a le nombre de gène dans son génome le plus petit du clade, soit 5916 gènes, mais si l'on rapporte le nombre de gènes en familles dynamiques au nombre total de gènes dans les génomes on observe que YAPH à tout de même la proportion la plus faible pour le clade (Tableau 3.4).

Tableau 3.4 : Le nombre de gènes en famille dynamique, leur proportion dans le génome, le nombre de gènes du génome et la taille de ce génome.

Espèces	Nb de gènes du génome (dont pseudo)	Taille du génome	Nb gènes en famille dynamique	‰ gènes en famille dynamique dans le génome
YALI	6642 (141)	20,5	85	12,80
YAGA	6936 (153)	22,8	82	11,82
YAYA	6283 (70)	18,8	70	10,14
YAPH	5916 (148)	16,2	58	9,80
YAAL	6209 (128)	19,8	80	12,88
OLHI	5338 (64)	10,7	44	8,24

Les gènes de ces familles dynamiques représentent entre 10,1 et 12,8‰ des gènes des génomes complets chez YALI, YAGA, YAYA et YAAL. Par contre chez YAPH ils ne représentent que 9,8‰ du total des gènes. En effet, avec ses 24 pertes branche spécifique, YAPH est l'espèce qui a perdu le plus de gènes depuis l'expansion qui a eu lieu après avant la divergence de YAAL (Figure 3.20). YAYA à subit 11 pertes, YAGA et YAAL deux et YALI une seule. Avec ses 9 duplications, YAPH est aussi l'espèce qui a subi le plus de duplications espèce spécifique mais ceci est biaisé par les 5 duplications successives du gène LIP2 (voir paragraphe 3.3.3). Ceci suggère tout de même que les familles impliquées dans le métabolisme des lipides sont plus particulièrement dynamiques chez YAPH avec une tendance à perdre des gènes. Ce phénomène s'observe aussi chez YAYA mais de manière plus modérée avec 11 pertes et 4 duplications spécifiques. Chez YAGA les duplications et les pertes spécifiques sont au nombre de 2. YALI et YAAL ont subi plus de duplications que de pertes de gènes depuis la divergence de YAAL.

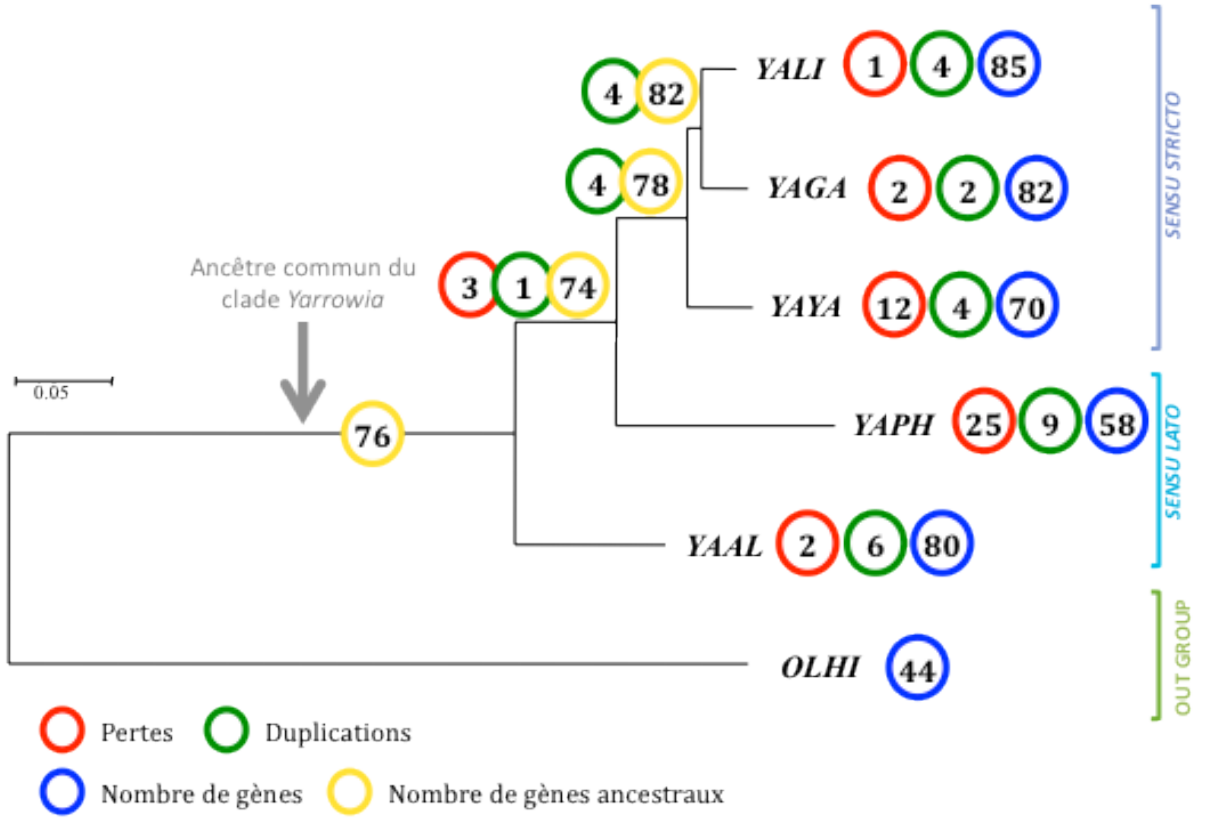


Figure 3.20 : Scénario évolutif des gènes en familles dynamiques dans le clade *Yarrowia*. Les évènements sont placés sur l'arbre phylogénétique des espèces construit à partir de la concaténation des 912 gènes (voir encadré page 73). Les duplications, pertes, le nombre de gènes de chaque espèce et le nombre de gènes ancestraux sont signifiés dans les ronds verts, rouges, bleus et jaunes, respectivement.

Les pseudogénisations sont observables dans 12 cas dont quatre cas au sein de la famille des lipases. Parmi ces 12 pseudogènes 4 sont très dégénérés et n'ont pas été inclus dans les phylogénies des familles. Avec le gène YyLIP5b de YAYA, nous pouvons observer un exemple de duplication génique suivi par la pseudogénisation de la nouvelle copie LIP5b mais l'arbre des LIP5 du clade, avec les deux LIP5 de YAGA qui ne branchent pas ensemble, est mal résolu (Figure 3.21). De même la duplication de SLC1 chez YAPH est suivie de la pseudogénisation de YpSLC1b. Là encore l'arbre phylogénétique est surprenant car SLC1 de OLHI se place entre YAPH et YAAL. Ceci pourrait s'expliquer par une divergence plus importante des SLC1 de YAPH. YpSLC1 est la copie ancestrale puisqu'elle est en synténie avec les autre SLC1 du clade et elle partage entre 73% et 82% avec les autres SLC1 ancestrales du clade. Par contre YpSLC1b est très divergente avec au maximum 33% de similarité partagée avec les autres SCL1 (Figure 3.22).

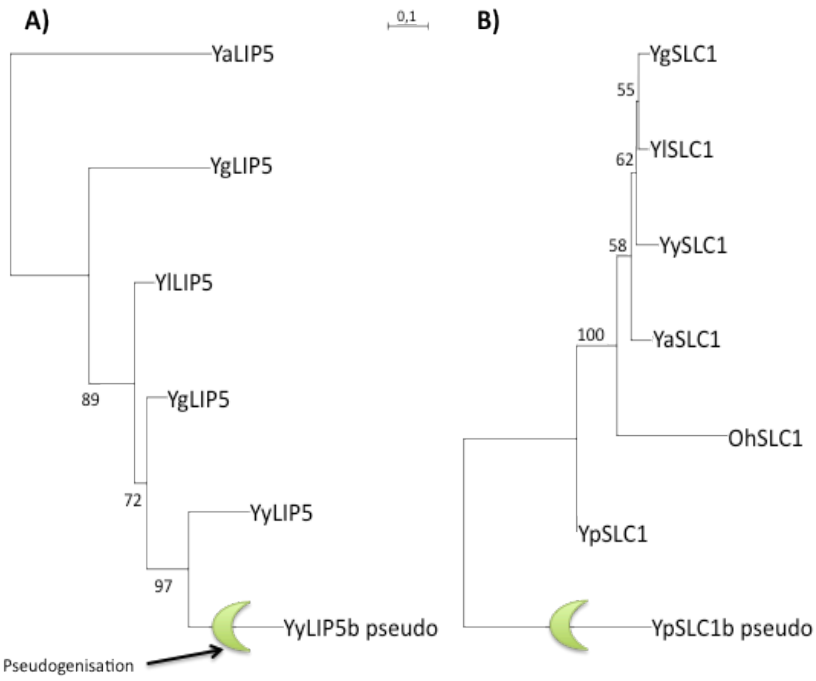


Figure 3.21 : A) Arbre phylogénique du gène LIP5 pour le clade *Yarrowia* et OLHI provenant d'un alignement en acides aminés, 308 résidus. B) Arbre phylogénique du gène SLC1 pour le clade *Yarrowia* et OLHI provenant d'un alignement en acides aminés, 338 résidus. Le symbole vert représente la pseudogénisation des gènes LIP5b chez YAYA et SLC1b chez YAPH.

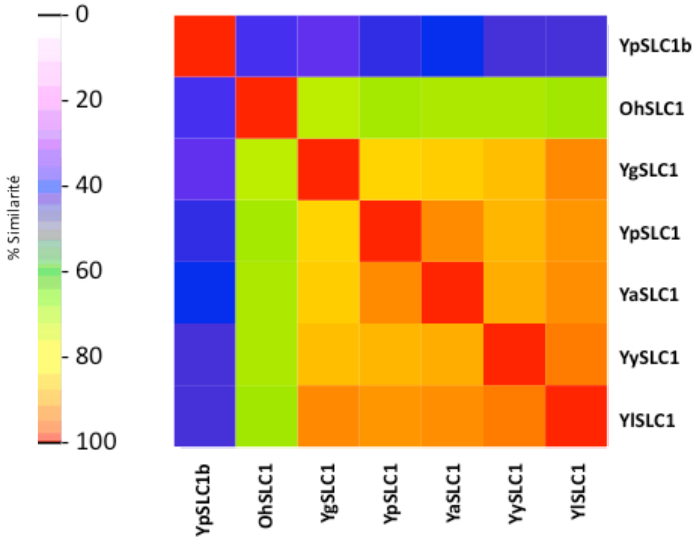


Figure 3.22 : Heatmap des pourcentages de similarité partagés entre les protéines SLC1 du clade *Yarrowia* et OLHI.

Sur les 12 cas de pseudogénisation, 5 sont des gènes de YAYA. Ceci va dans le même sens que ce que nous avons observé précédemment, c'est à dire que YAYA, comme YAPH, semble subir une contraction des familles de gènes impliquées dans le métabolisme lipidique.

2.3 Des familles sous pression

La moyenne des dN/dS des paires de gènes de familles gelées est plus faible que celle des paires de gènes des familles dynamiques, $dN/dS_{moyenGelées}$=0,053 pour les gelées et $dN/dS_{moyenDyn}$=0,068 pour les dynamiques. Les pressions de sélection qui s'exercent sur les familles gelées et sur les

familles dynamiques sont dans les deux cas des forces purifiantes. Elles sont comparables pour les familles dynamiques et le reste des génomes ($dN/dS_{moyenGénome}$=0,072) mais elles semblent agir plus fortement sur les gènes des familles gelées (Figure 3.23). Les moyennes des ratios de similarité entre ces deux types de familles sont également différentes, 0,97 pour les dynamiques et 1,0 pour les gelées. Toutes les familles gelées, sauf TGL, et les sept familles dynamiques ABC, CNT2, GCY, LSC2, PDC1, POX, SPS ont des ratios de similarité supérieurs à 1 et semblent donc avoir des gènes plus contraints que le reste des génomes du clade. Par contre les familles très dynamiques, LIP, HFD, EST, SLC1 et CNT4, et les familles peu dynamiques, ADH, ALK, COU et ISO, ont des ratios inférieurs à 1 ce qui suggère qu'elles contiennent des gènes peu contraints. Ce résultat suggère que dans les familles ayant subi des duplications, certaines copies de gènes dupliquées sont moins soumises à pression purifiante que les gènes présents en une seule copie dans un génome.

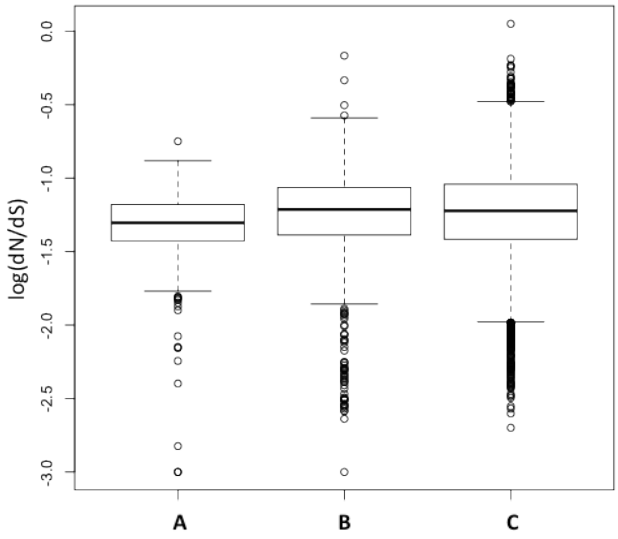


Figure 3.23 : Boxplot des dN/dS, représentés en log pour : A) les gènes des familles gelées, B) les gènes des familles dynamiques, C) tous les orthologues synténiques qui ne sont pas impliqués dans le métabolisme lipidique.

3 Et plus précisément ?

Nous avons décidé d'étudier plus particulièrement deux exemples de famille peu dynamique, les gènes POX dont ont connaît bien leur fonction respective et la famille EST avec un Dyn égal à 1 et un exemple de famille très dynamique, les gènes LIP très étudiés pour le métabolisme lipidique.

3.1 POX, un exemple de famille peu dynamique

Les gènes POX codent pour des acyl-CoA oxydases essentielles pour la dégradation des acides gras puisqu'elles réalisent la première étape de la ß-oxydation dans les peroxysomes.

La famille POX, avec un Dyn de 0,5, est la famille la moins dynamique. L'ancêtre commun avait 6 gènes POX et chaque génome du clade en a également six (Figure 3.24). Dans le cas de cette famille, l'expansion du nombre de gènes s'est fait avant la divergence de OLHI qui a aussi 6 gènes POX. Les seuls évènements observables depuis la divergence de YAAL sont la perte du gène POX4 chez YAPH, la duplication du gène POX5 chez la même espèce et la duplication d'une seconde copie de POX1 chez YAGA. Nous pouvons également observer sur l'arbre phylogénétique de la famille POX (figure 3.24) la forte divergence du gène POX4 de OLHI.

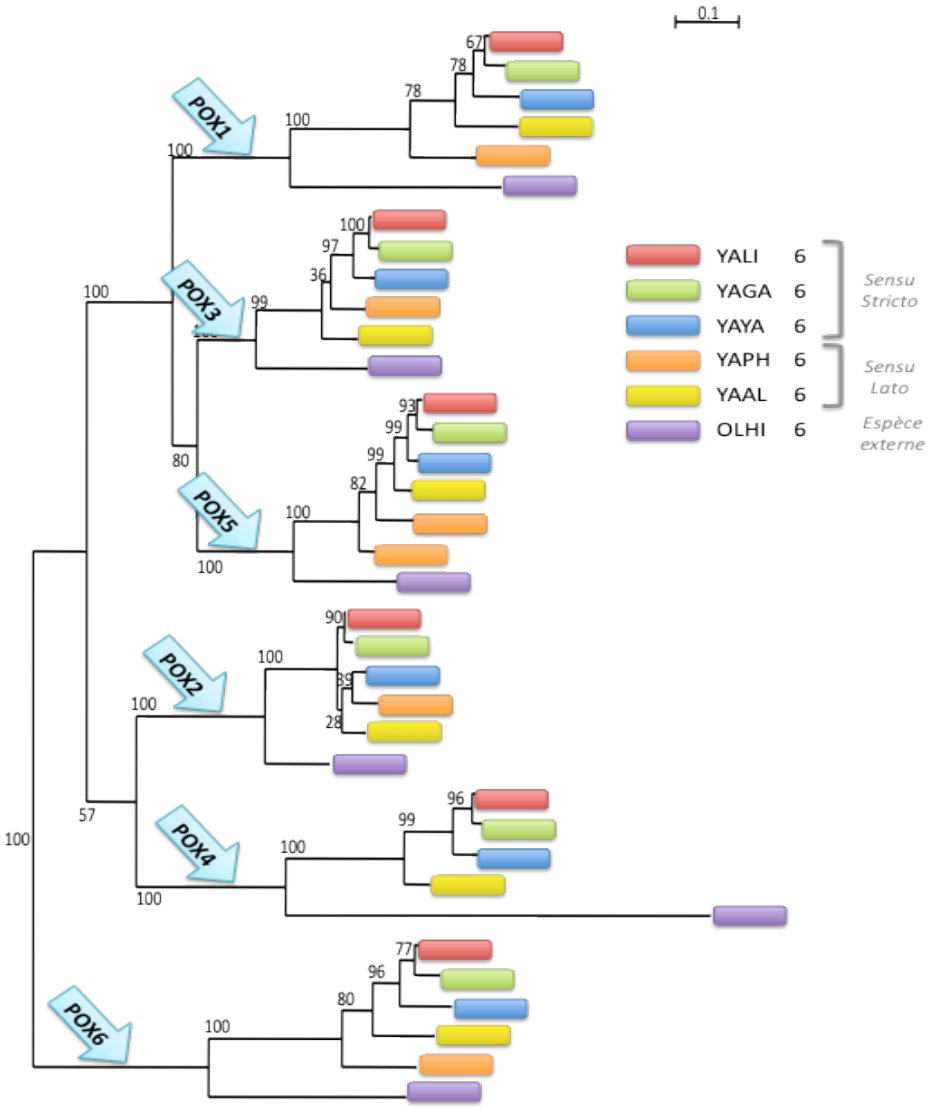


Figure 3.24 : Phylogénie de la famille POX pour le clade *Yarrowia*. En bas à droite : le code couleur et le nombre de gènes pour chaque espèce.

Chez les *sensu stricto*, tous les gènes POX sont en synténie. La région contenant POX4 chez YALI est en synténie avec une région du scaffold 2 de YAPH. Dans cette région chez YAPH on observe la perte d'un gène sur deux en aval de la position où POX4 a été perdu (Figure 3.25A). Le gène POX5 a été dupliqué chez YAPH. Cette duplication ne s'est pas faite en tandem (Figure 3.25B et C). La copie ancestrale est en synténie avec POX5 de YALI avec cependant une rupture de

synténie directement en aval de YpPOX5. L'autre copie est dans une région très remaniée du scaffold 4 qui présente encore une synténie avec une région du chromosome A de YALI.

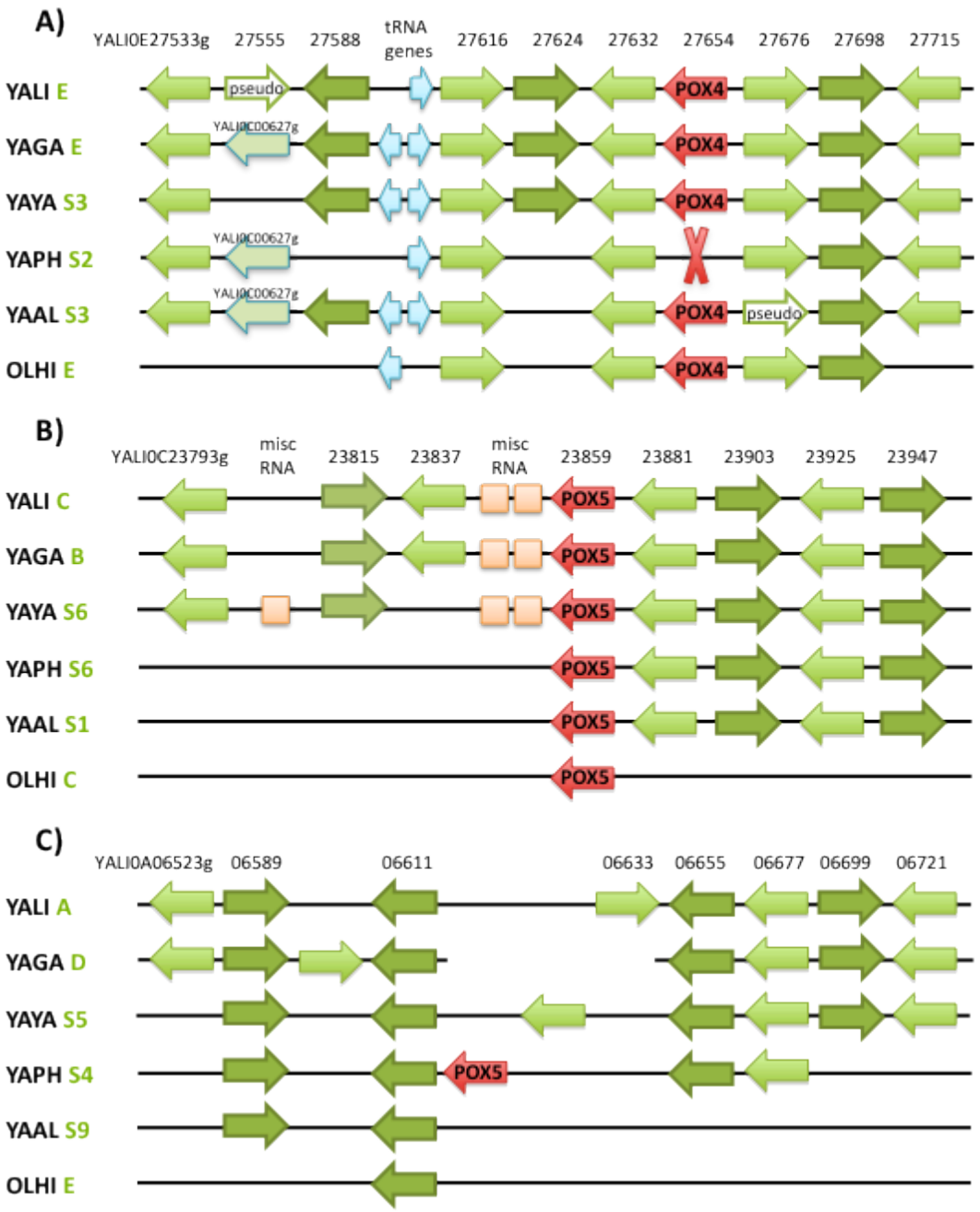


Figure 3.25 : A) Représentation des gènes homologues dans la zone contenant le gène POX4 dans le clade *Yarrowia*. B) Représentation des gènes homologues dans la zone contenant le gène POX5 dans le clade *Yarrowia*. C) Représentation des gènes homologues du clade *Yarrowia* dans la zone contenant le gène POX5b chez YAPH.

De manière générale chez YAPH tous les POX, sauf POX5b, sont en synténie par rapport à YALI mais pour POX1, POX3, POX5a et POX6 il y a une rupture de synténie directement en amont ou en aval du gène (voir par exemple POX1 en Figure 3.26).

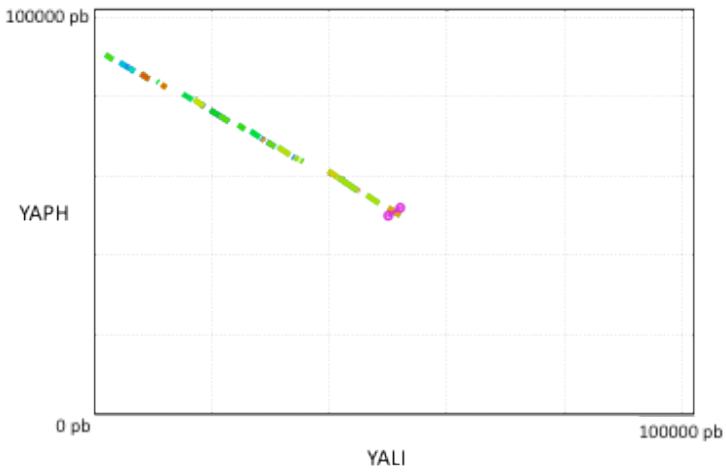


Figure 3.26 : Exemple d'une rupture de synténie directement après un gène POX. Dotplot de YAAL contre YALI du gène POX1 avec 50000 pb en amont et en aval. Le gène POX1 est représenté en violet.

Chez YAAL, tous les POX sont également en synténie par rapport à YALI mais nous observons le même type de rupture de synténie pour POX1, POX2, POX5 et POX6.

L'organisation dans les régions de POX1 et POX5 de YAAL et YAPH est ancestrale car elles sont parfaitement en synténie (Figure 3.27A et B). La rupture de synténie s'est alors faite juste avant la divergence des *sensu stricto* pour ces deux gènes. Chez les trois espèces *sensu stricto*, directement en aval du gène, là où il y a eu rupture de synténie, il y a une ou deux séquences non codantes transcrites annotées en misc-RNA. Ces misc-RNA de YALI et YAYA contiennent deux paires de séquences orthologues, l'une de 408 bases et l'autre de 258 bases.

La première paire s'aligne avec le gène homologue YALI0B23496g de 136 aa (e-values de 6^{e}-25 et 9^{e}-15 respectivement). Les séquences de l'autre paire s'alignent avec un gène de YALI, YALI0A21351g, avec des e-values de 6^{e}-10 pour YAYA et 1^{e}-39 pour YALI.

Par contre les misc-RNA de YAGA n'ont aucune séquence commune avec ceux de YALI et YAYA. Une première partie d'un misc-RNA de YAGA s'aligne avec un gène de YAGA, YAGA0D00137g, de 142 aa (e-value = 2^{e}-7) et une seconde partie avec un gène de YALI, YALI0A06633g, de 164 aa (e-value de 6^{e}-9). Tous ces gènes sont annotés comme sans similarité et probables séquences non codantes.

Pour POX6 la synténie n'est pas la conservée entre YAAL et YAPH. Chez YAAL il y a rupture de synténie par rapport à YALI directement en amont du gène alors que chez YAPH elle est directement en aval (Figure 3.27C). Les ruptures de synténie au niveau des POX6 semblent donc espèce spécifique. Enfin chez OLHI seul POX4 est en synténie avec YALI (Fig a). Tous les autres gènes POX sont dans des régions très remaniées.

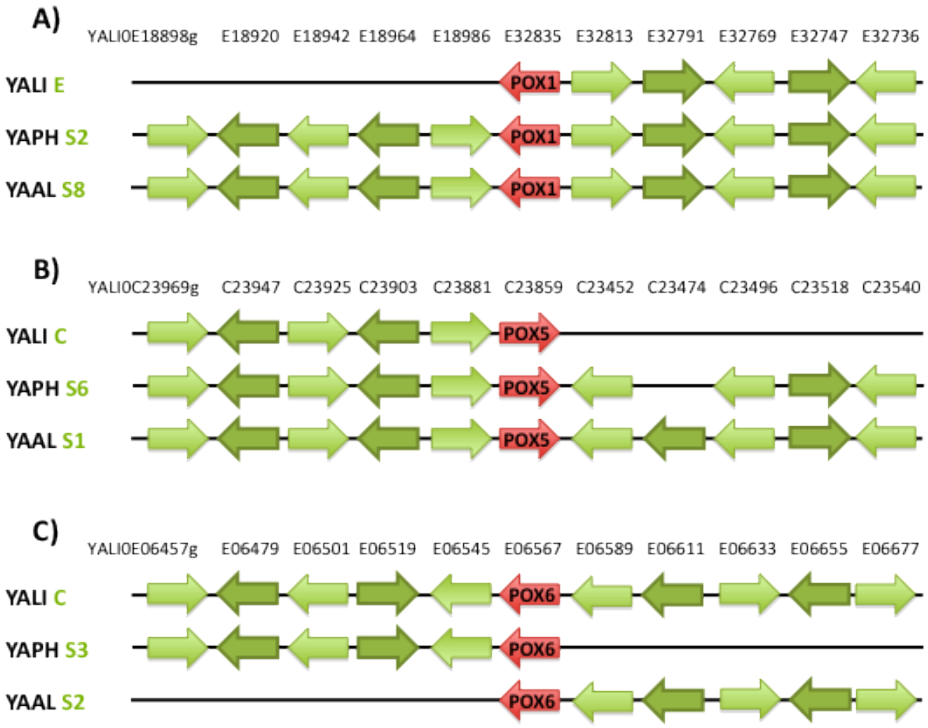


Figure 3.27 : A) Représentation des gènes homologues dans la zone contenant le gène POX1 dans YAAL, YAPH et YALI. B) Représentation des gènes homologues dans la zone contenant le gène POX5 dans YAAL, YAPH et YALI. C) Représentation des gènes homologues dans la zone contenant le gène POX6 dans YAAL, YAPH et YALI.

Au sein du clade les gènes POX sont très proches, avec des similarités moyennes pour le clade comprises entre 93,2 % pour POX6 et 96,8 % pour POX2 (Figure 3.28A). La moyenne des ratios de similarité des gènes POX, toujours supérieure à 1, témoigne de la contrainte s'exerçant sur ces gènes (Figure 3.28B). La similarité la plus élevée, de 99,6%, est observée pour les gènes POX2 pour le couple YALI-YAGA, et la plus faible est celle pour le gène POX1 entre YALI et YAPH.

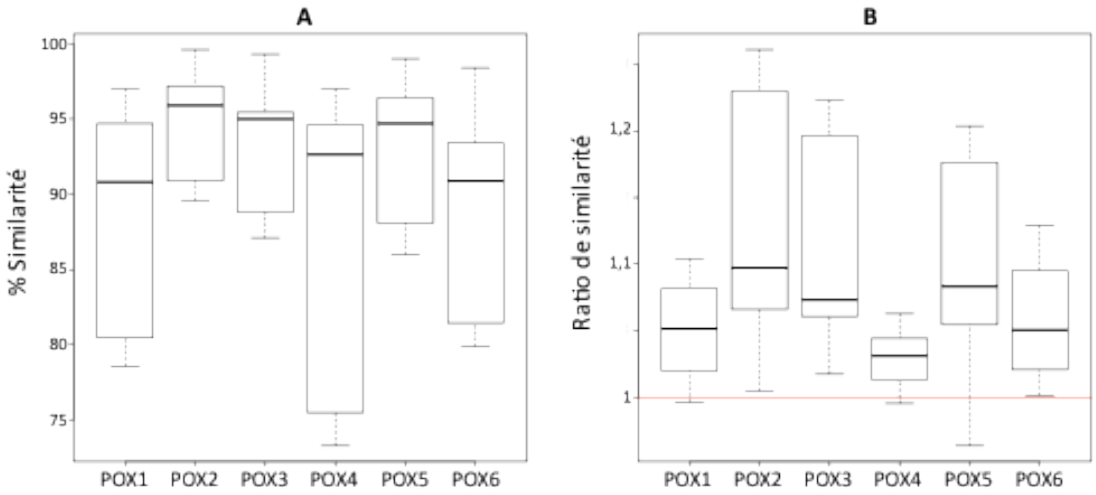


Figure 3.28 : A) Boxplots des pourcentages de similarité entre les gènes POX du clade *Yarrowia*. B) Boxplot des ratios de similarité entre les gènes POX du clade *Yarrowia*. Un ratio égal à 1 est marqué par la ligne rouge.

Les six gènes POX ont des patterns d'expression similaires pour les espèces YALI, YAGA et YAPH. POX1, POX4 et POX6 sont peu, ou pas, exprimés sur glucose, acide oléique et tributyrine (Tableau 3.5). En revanche POX2, POX3 et POX5 sont très fortement exprimés sur acide oléique. La seconde copie de POX4 chez YAPH a sur ces trois milieux les mêmes niveaux d'expression que la copie ancestrale. Elle est également très fortement exprimée en acide oléique. Nous pouvons donc penser que cette duplication confère à YAPH un avantage pour pousser sur milieux hydrophobes en optimisant sa ß-oxydation.

Tableau 3.5 : Expression relative des gènes de la famille POX sur les milieux glucose, acide oléique et tributyrine. Les données proviennent des transcriptomes déposés à l'ENA-EMBL sous le numéro de projet PRJEB6632 (voir article page 163). La seconde copie du gène POX5 de YAPH est en rouge.

		Glucose	Acide oléique	Tributyrine
POX1	YALI	+/-	+/-	+/-
	YAGA	-	-	-
	YAPH	-	-	-
POX2	YALI	++++	+/-	+/-
	YAGA	+++	+/-	+/-
	YAPH	++++	+	-
POX3	YALI	++++	+/-	+/-
	YAGA	+++	++	+
	YAPH	+++	++	++
POX4	YALI	+/-	-	-
	YAGA	+/-	-	-
POX5	YALI	++++++	+/-	-
	YAGA	++++	+/-	-
	YAPH	++++	+/-	-
	YAPH b	++++	+/-	-
POX6	YALI	+/-	+/-	-
	YAGA	+	+/-	+/-
	YAPH	+	+/-	-

3.2 Les EST, un cas d'école

Les gènes de la famille EST codent des estérases avec une fonction hydrolase ou lipase. Elles permettent de couper un ester en un acide et un alcool. Cette famille à un Dyn = 1, elle a donc subit autant d'évènements qu'elle a de gènes ancestraux. L'ancêtre commun au clade avait 4 gènes EST, YALI en a 4, YAGA 3, YAYA 5, YAPH 2, YAAL 4 et OLHI n'en a qu'un seul (Figure 3.29). La synténie des quatre gènes de la famille est conservée dans toutes les espèces du clade ce qui permet de retracer aisément les évènements qui ont affectés cette famille. En reconstruisant le scénario évolutif nous avons compté une perte chez YAGA, deux chez YAPH et une duplication chez YAYA (Figure 3.30).

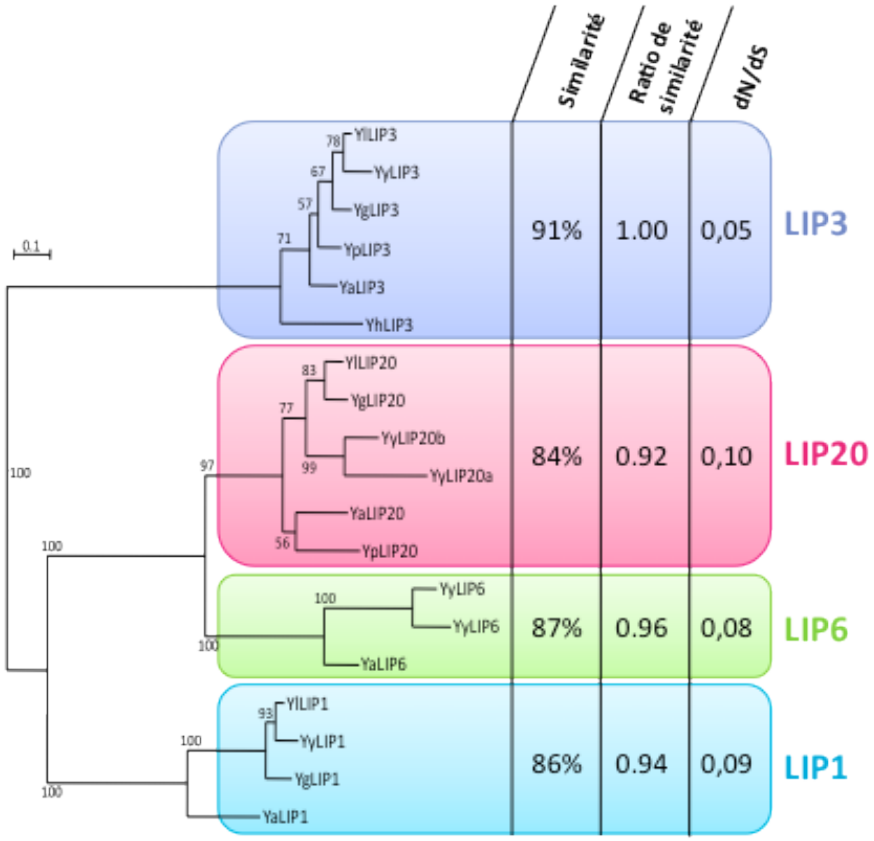


Figure 3.29 : Phylogénie de la famille EST pour le clade *Yarrowia* et OLHI, construit à partir de la concaténation de 263 résidus.

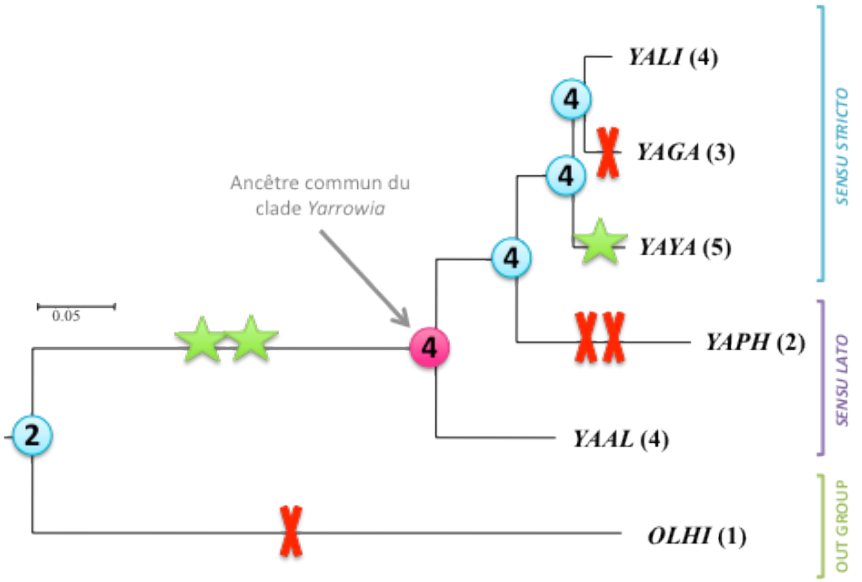


Figure 3.30 : Scénario évolutif des gènes de la famille EST dans le clade *Yarrowia* et OLHI. Les évènements sont placés sur l'arbre phylogénétique des espèces construit à partir de la concaténation de 912 gènes (voir encadré page 73). Les étoiles vertes et les croix rouges symbolisent respectivement les duplications et les pertes. Le nombre de gènes pour chaque espèce est donné entre parenthèse et le nombre de gènes ancestraux est signifié dans les ronds colorés.

L'ancêtre commun à OLHI et au clade *Yarrowia* a deux gènes EST. A l'image de la majorité des familles dynamiques impliquées dans le métabolisme lipidique, l'expansion a principalement eu lieu après la divergence de OLHI avec deux duplications.
Nous pouvons observer dans la phylogénie de la famille EST qu'une première LIP, commune au clade et à OLHI, s'est dupliquée pour donner LIP3 et une autre LIP. LIP3, qui est alors une des deux copies ancestrales au clade et à OLHI, n'a subi aucun événement au cours l'évolution du clade et de OLHI. L'ancêtre des LIP20, LIP6 et LIP1 par contre à été perdu spécifiquement chez OLHI et s'est dupliqué consécutivement deux fois après la divergence de OLHI, pour donner d'abords LIP1 puis LIP20 et LIP6.
Les moyennes des pourcentages de similarité et des ratios de similarité pour chaque paralogue montrent que la copie ancestrale LIP3 tend à être plus conservée. En effet les LIP3 ont 91% de similarité entre elles et un ratio moyen de 1,02, alors que les autres EST partagent entre 84% et 87% de similarité et leurs moyennes des ratios de similarité sont inférieures à 1 (de 0,92 à 0,96). De même la moyenne des dN/dS de LIP3, de 0,05, est presque deux fois plus faible que celles des autres gènes de la famille, allant de 0,08 à 0,10.
Ceci suggère que les gènes descendant de la copie ancestrale LIP3 ont moins divergés que les autres gènes de cette famille et que ces gènes sont sous une pression purifiante plus forte. De façon générale ce gène LIP3 est plus contraint que les autres gènes des familles dynamiques ($dN/dS_{moyenDyn}$=0,068) et, à fortiori, plus contraint que l'ensemble des orthologues synténiques des génomes complets du clade ($dN/dS_{moyenGénome}$=0,072). Son dN/dS moyen est du même ordre que celui des familles gelées ($dN/dS_{moyenGelées}$=0,053) qui comme les LIP3 n'ont pas subi d'évènements évolutif depuis la divergence du clade *Yarrowia*.

3.3 LIP, un exemple de famille très dynamique

Les lipases sont une catégorie spécifique des estérases puisqu'elles sont spécialisées dans la transformation de triglycérides en glycérols et en acides gras. En retraçant le scénario évolutif de la famille des lipases, LIP, nous avons déterminé qu'il y avait 9 lipases ancestrales. Après 15 duplications et 11 pertes dans le clade il y a jusqu'à 16 LIP chez YALI et au minimum 9 LIP chez YAYA et YAPH. Chez OLHI on observe seulement 3 lipases dont deux proviennent d'une duplication probablement espèce spécifique. Entre OLHI et l'ancêtre commun au clade, il faut au moins 7 duplications pour expliquer la différence du nombre de copies.
La famille des LIP est alors un exemple de famille très dynamique, avec un ratio de 3,3 soit plus de trois évènements par gène ancestral.

L'étude de l'évolution de la famille LIP a donné lieu à l'écriture d'un article de recherche en collaboration avec Muchalin Meunchan, actuellement en doctorat sous la direction de Alain

Marty dans l'équipe Ingénierie des Systèmes Biologiques et des Procédés à LISBP (INRA, UMR792) à Toulouse. L'article, présenté à la suite de ce paragraphe, est en cours de finalisation et sera bientôt soumis. Leur laboratoire s'intéresse, depuis plusieurs années, à l'activité de YlLIP2 et plus particulièrement à son enantiosélectivité. Pour ce projet, nous avons aligné les différentes lipases, repéré les sites actifs et les sites d'interaction, et nous avons corrélé les changements d'acides aminés et les résidus importants avec des changements d'activité. Nous avons ensuite reconstruit le scénario évolutif de cette famille, étudié les expressions relatives des lipases de YALI, YAGA et YAPH, détecté un promoteur fort régulé par le milieu et caractérisé les pressions de sélection agissant sur les LIP2. Les activités enzymatiques et l'enantiosélectivité de toutes les LIP2 du clade ont ensuite été testées à LISPB dans différents mutants et variants.
Ce travail effectué sur la famille des LIP donne un exemple de ce qui peut être fait en créant des liens entre différentes approches.

Evolutionary genomics: an efficient tool to explore enzyme diversity and guide their engineering.

Muchalin Meunchan[1,2,3,4,#], Stéphanie Michely[5,6,#], Hugo Devillers[5,6], Jean-Marc Nicaud[5,6], Alain Marty[1,2,3,*], Cécile Neuvéglise[5,6,*]

[1] Université de Toulouse; INSA, UPS, INP; LISBP, 135 Avenue de Rangueil, F-31077 Toulouse, France

[2] INRA, UMR792 Ingénierie des Systèmes Biologiques et des Procédés, F-31400 Toulouse, France

[3] CNRS, UMR5504, F-31400 Toulouse, France

[4] Department of Biochemistry, Faculty of Science, Khon Kaen University, 123 Mittapap Road, Khon Kaen, 40002, Thailand

[5] INRA, UMR 1319 Micalis, F-78352 Jouy-en-Josas, France

[6] AgroParisTech, UMR Micalis, F-78352 Jouy-en-Josas, France

# These two authors contributed equally to this work.

* Corresponding authors
Cécile Neuvéglise:
Tel (+33) 1 30815478, Fax (+33) 1 30815457, Email ncecile@grignon.inra.fr
Alain Marty:
Tel (+33) 5 61559439, Fax (+33) 5 61559400, Email alain.marty@insa-toulouse.fr

Key-words

Yarrowia lipolytica, lipase, enantioselectivity, oleaginous yeast, genome evolution, gene family

Abstract

To promote the industrial use of enzymes, it is essential to engineer enzymes that meet industrial specifications, particularly in terms of selectivity. Success means identifying the most promising enzyme as a target for evolution. Here, we propose an original way to discover new successful enzymes. The idea is to use evolutionary biology to search for homologues of an identified enzyme of interest in closely related sequenced species.

Lipases are currently the subject of intensive studies due to their large range of industrial applications. The Lip2p lipase from the yeast *Yarrowia lipolytica* (Yl*LIP2*) was recently shown to be a good candidate for different biotechnological applications. We showed it to be an efficient catalyst for the resolution of 2-bromo-arylacetic acid esters, an important class of chemical intermediates in the pharmaceutical industry. Using a combination of comparative genomics approaches based on sequence similarity, synteny conservation, and phylogeny, we constructed the evolutionary scenario of the lipase family for six species of the *Yarrowia* clade. RNA-seq based transcriptome analysis revealed the primary role of LIP2 homologues in the assimilation of different substrates. Once identified, these Yl*LIP2* homologues were expressed in *Y. lipolytica*. The lipase Lip2a from *Candida phangngensis* was shown to naturally present better activity and enantioselectivity than YlLip2. Enantioselectivity was further improved by site-directed mutagenesis targeted to the substrate binding site. The mono-substituted variant V232S displayed enantioselectivity greater than 200 and a 2.5 fold increase in velocity. A double-substituted variant 97A-V232F presented reversed enantioselectivity, with a total preference for the *R*-enantiomer.

Introduction

Lipases are serine hydrolases defined as triacylglycerol acylhydrolases (E.C. 3.1.1.3). They are ubiquitous enzymes of high physiological significance and their industrial potential is on the increase. Their physiological role is to catalyse the hydrolysis of the ester bond of tri-, di- and monoglycerides of long-chain fatty acids into fatty acids and glycerol (Schrag and Cygler 1997). Lipases are widely used as additives in industrial laundry and household detergents (Cardenas et al. 2001), in the food industry for, for instance, the development of flavours, the production of structured lipids or the production of oil enriched in polyunsaturated acids, paper manufacture, degradation of fatty waste (Masse et al. 2001), and in the synthesis of fine chemicals, cosmetics and pharmaceuticals.

Fungi are known to secrete lipases to assimilate lipid substrates from their natural environment. Being ubiquitous, lipases have developed substrate specificities and stability under a range of chemical and physical conditions, which increases their interest for industrial applications. In addition, some fungal lipases are extracellular, thus reducing the cost of production and making this source preferable over bacteria. Among fungal lipases (Gopinath et al. 2013), those of ascomycetous yeasts from various species such as *Candida rugosa* (Benjamin and Pandey 1998), *Yarrowia* (*Candida*) *deformans* (Bigey et al. 2003), *Candida albicans* (Hube et al. 2000), *Candida viswanathii* (de Almeida et al. 2013), *Pseudozyma* (*Candida*) *antarctica* (Tan et al. 2010) or *Yarrowia lipolytica* (Fickers et al. 2011) have been widely studied. The lipases in most of these species belong to multigenic families. Up to 16 members have been reported in *Y. lipolytica*, but only the extracellular YlLip2 lipase has been extensively studied (Pignede et al. 2000), and was reported to be an efficient enzyme in a number of applications (Fickers et al. 2005). For instance, it was demonstrated to be the most effective lipase for the purification of *cis*-4,7,10,13,16,19-docosahexaenoic acid (DHA) by hydrolysis of a mixture of ethyl esters from tuna oil (Casas-Godoy et al. 2014) and to be an efficient stereo-selective enzyme for the resolution of 2-halogeno-arylacetic acid esters, an important class of chemical intermediates in the pharmaceutical industry (Guieysse et al. 2004).

Studies have been conducted to improve YlLip2 properties, e.g. lipase activity, enantioselectivity, and thermostability. The strategies used were mainly based on site-directed mutagenesis targeted to the substrate binding site (Bordes et al. 2009) or on random mutagenesis of the whole gene sequence (Bordes et al. 2011). For instance, the mono-substituted variant with the valine 232 changed into serine in YlLip2, represents a tremendous increase in enantioselective activity compared to the parental enzyme for the resolution of 2-bromo-phenylacetic acid ethyl ester (58-fold) and 2-bromo-o-tolylacetic acid ethyl ester (16-

fold) (Bordes et al. 2009). This type of experimental approach requires the screening of a large set of mutants and is thus laborious and, in addition, the outcome is uncertain.

In this paper, we propose an original approach to improve enzymes and their evolution. The approach used takes advantage of the decreasing cost of genome sequencing, thereby enabling a huge number of closely related organisms to be sequenced. *Y. lipolytica* is an interesting reservoir of lipases of which only few members have been exploited to date. These paralogue lipases, which derived from gene duplications, are already too divergent to deduce relationships between activities and amino-acid sequences. An alternative approach is to identify more conserved lipases derived from a common ancestor, which could have acquired improved specificities. Combining comparative genomics and phenotypic characterisation may provide clues to interesting mutations to improve lipase specificities of high interest for industry. From this perspective, we investigated the genomes of six members of the *Yarrowia* clade and detected 61 lipase genes belonging to different LIP families. The study of the evolution of the lipases made it possible to identify the orthologous relationships between homologues and to retrace their evolutionary history. With 11 members closely related to YlLip2, the Lip2 family is the largest and the most dynamic one. We consequently investigated this family and tried to infer the specificities of each member by over-expressing them in *Y. lipolytica* under the control of the strong constitutive TEF promoter and to identify the best lipase for resolution of 2-halogeno-arylacetic acid esters.

Material and Methods

Strains

The strains of the *Yarrowia* clade investigated in this study are listed in Table 1. In addition, *Candida hispaniensis* was used as the nearest out-group species for the *Yarrowia* clade (Michely et al., in prep). All were wild-type prototroph strains. Strain names are abbreviated as follows: YALI (*Y. lipolytica* W29=CBS 7504), YAYA (*Y. yakushimensis* CBS 10253), YADE *(Y. deformans* CBS 2071), YAGA *(Y. galli* CBS 9722), YAOS *(Y. oslonensis* CBS 10146), YAHO *(Y. hollandica* CBS 4855), YAPH *(Y. phangngensis* CBS 10407), YAAL *(Y. alimentaria* CBS 10151), CAHI *(C. hispaniensis* CBS 9996).

Identification of lipases in species of the Yarrowia clade

The genomes of YAGA, YAYA, YAPH, YAAL, and CAHI were sequenced and annotated at INRA Thiverval-Grignon. Genome annotations were performed by a combination of *in silico* annotation transfer tool based on Amadea BioPack (ISoft, France) and manual curation based on RNA-Seq data. Genes encoding lipases (LIP genes) were found by homology using the LIP genes of YALI (Edgar 2004). Orthologues in each species were identified based on a two-step reciprocal approach using BLASTP. In the first step, YALI LIP genes were used as queries for BLASTP on the 5 other species, with a cutoff E-value of $1.e^{-10}$ (Altschul et al. 1990). Then a reciprocal BLASTP search considering only the best hit allowed genes corresponding to other YALI genes to be discarded. The sequence of all LIP genes have been deposited at ENA-EMBL under accession numbers XX to XX. Their protein sequences are provided in File S1.

RNA-Seq analysis

YALI, YAGA and YAPH were used for transcriptome analysis in three different media, i.e. glucose (GL), oleic acid (OA) and tributyrin (TB). The strains were cultured at 28 °C on YPD medium (1% (wt/vol) yeast extract, 1% (wt/vol) peptone, 1% (wt/vol) glucose) for precultures. Then, minimal medium base (MMB) was used, containing 0.17% (wt/vol) yeast nitrogen base without amino acids and ammonium sulfate (Difco, Paris, France), 0.5% (wt/vol) NH_4Cl, 50 mM phosphate buffer pH 6.8 and 0.15% (wt/vol). The three carbon sources (GL, AO, TB) were added at a final concentration of 1%. Both lipid substrates were previously emulsified by sonication of a 20% mixture in the presence of 0.625% Tween 40. Cells grown in the presence of OA or TB were washed twice with 0.5% bovine serum albumin and then once with 0.9% NaCl before OD_{600} determination. Cells were harvested in the exponential phase. Two replicates were performed for each species/media pair. Total

RNA was prepared using the Qiagen RNeasy kit (Qiagen, Courtaboeuf, France). mRNA was purified by selection of poly(A)+ transcripts, which were then sequenced by the Illumina Solexa technology with a HiSeq 2000 sequencing system with chemistry v3.0. Twelve to 29 million single-end reads of 100 bp were generated per sample. Sequences were cleaned and trimmed using Trimmomatic version 0.20 (http://www.usadellab.org/cms/index.php?page=trimmomatic) with arguments -threads 20 and -phred33. Only reads with a minimum length of 40 bp, which corresponded to about 98% of the raw data, were further analysed. Tophat2 v2.0.10 with Bowtie2 v2.1.0 were used to map the reads on the nuclear genome of the three strains (Kim et al. 2013). The options used were the following: --micro-exon-search --min-intron-length 30 --min-coverage-intron 30 --min-segment-intron 30 --max-intron-length 4000 --max-multihits 1 with the transcriptome being provided. Reads were counted using a custom BioPerl script based on samtools utilities v0.1.18 (Li et al. 2009). In a preliminary step, low expression features were filtered as they generally cause serious biases in differential expression studies. To this end, we followed the procedure described in the egdeR user guide (Robinson et al. 2010). Thus, only features with at least one count per million (cpm) in at least two conditions/replicates were kept. Differential expression was performed with DEseq2 v1.2.8 (Anders and Huber 2010). The default normalization proposed by the DEseq2 package was used and the "parametric" estimation of data dispersion was chosen. All the pairwise comparisons of the three different conditions for the three species were investigated. The resulting p-values were corrected using the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995). Assuming that the variability between replicates was lower in YALI than in YAGA and YAPH (data not shown), the adjusted P-value cut-off for the differential expression was fixed at 1.e-3 in YALI and 5.e-3 in YAGA and YAPH. Fastq files of the RNA-Seq reads have been deposited at the European Nucleotide Archive (ENA-EMBL) under the project number PRJEB6632 (http://www.ebi.ac.uk/ena/data/view/ERP006181).

Reconstruction of the evolutionary scenario

To reconstruct the evolutionary scenario of the lipase family in the *Yarrowia* clade using CAHI as out-group, the species tree and the LIP tree were reconciled with synteny data. A parsimonious approach was used to minimise the number of duplications and loss of LIP genes. For the phylogeny of the LIP genes, protein sequences were aligned with MAFFT (Katoh et al. 2002). The pairwise comparisons of both amino acid identity and similarity were calculated using aligncopypair (EMBOSS) (Rice et al. 2000).

The phylogenetic tree deduced from the LIP protein alignment was constructed with PhyML v3.0 with a JTT substitution model corrected for heterogeneity among sites by a Γ-law distribution using 4 different categories of evolution rates (Guindon and Gascuel 2003). The proportion of invariable sites and the α-parameter of the Γ-law distribution were optimised according to the data. A bootstrap value was calculated with 100 replicates. The species tree was deduced from the alignment of 912 single copy protein-coding genes (398959 residues). Individual gene alignments for the 912 orthologues was performed with MUSCLE (Edgar 2004) edited using Gblocks (Castresana 2000), concatenated, and the phylogenetic tree was estimated by maximum likelihood using PhyML v3.0 assuming a JTT substitution model with Γ distributed rate variation and a proportion of invariant sites estimated from the data.
Synteny conservation between LIP genes was determined by pair-wise comparison between 50 kbp regions upstream and downstream from each LIP gene, using Promer and mummerplot from MUMmer v3.23 (Kurtz et al. 2004).

Lip2 protein sequence analysis

Lip2 proteins were aligned with Multalin (Corpet 1988). Pairwise identity and similarity between Lip2 proteins were deduced from this alignment using aligncopypair (EMBOSS) (Rice et al. 2000). The secondary structure of the Lip2 proteins was predicted with PSIPRED v3.3 (Buchan et al. 2013). β-sheets were numbered according to the nomenclature of (Ollis et al. 1992) and the α/β hydrolase fold according to (Bordes et al. 2010).

Selection and dN/dS ratio

Pairwise nucleotide alignment was deduced from MAFFT protein alignment using tranalign in the EMBOSS package (Rice et al. 2000). The dN/dS ratios were calculated from this alignment using Codeml model from PAML4 package version 4.4b with a run mode equal to -2 (Yang 2007). dS values higher than 2.5 were removed from the analysis to avoid dS saturation, which notably removed all pairwise comparisons involving CAHI. Box plots were constructed with R (R Development Core Team 2011).

Chemical reagents

Yeast extract, tryptone, and peptone were purchased from Difco (Paris, France). All reagents used for enzymatic reactions were of commercial quality and were purchased from Sigma-Aldrich (St. Louis, MO, US). *n*-Decane was dried over molecular sieve (3 Å) before use. The preparation procedure for (±) 2-bromo-phenylacetic acid octyl ester is described in a previous paper (Guieysse et al. 2003).

Construction of mutants for production of recombinant enzymes

Plasmid JMP62-*URA3*ex-pTEF was used for the production of recombinant enzymes, wild-type and variants (Piamtongkam et al. 2011). This plasmid carries the wild-type/mutated *LIP2* genes under the control of the constitutive TEF promoter (pTEF). The mature sequence of wild type Lip2 from species of the *Yarrowia* clade and CAHI were integrated downstream either the PrePro-Lip2 secretion signal of *Y. lipolytica* or their own PrePro secretion signal.

For constructions using the PrePro-Lip2 secretion signal of *Y. lipolytica*, forward and reverse primers were designed to amplify the mature sequence of each Lip2 gene from species of the *Yarrowia* clade and CAHI, and to introduce *Bsrg*I and *Avr*II restriction sites. The PCR product obtained from genomic DNA was digested with these restriction enzymes and ligated to the plasmid JMP62-*URA3*ex-pTEF digested with the same restriction enzymes and gel purified. For constructions using their own PrePro signal secretion, the In-Fusion HD Cloning Plus kit from Clontech (Mountain View, CA, US) was used. Primers were designed to amplify the complete sequence of each Lip2 gene, including the PrePro secretion signal, and at each extremity, to introduce 15 bases homologous to those of the vector linearized by PCR amplification using both primers aJMP62TEFPPf and aJMP62TEFLip2YLr.

Variants V232S and D97A-V232F of YpLip2a were constructed using site directed mutagenesis. PCR amplification was performed using plasmid JMP62-YpLip2a as template and overlapping primers containing the desired mutation. PCR reaction was subsequently digested by *Dpn*I and directly transformed in *E. coli* DH5α strain.

All primers were purchased from Eurogentec (Angers, France) and the sequences are listed in Table S3. For all constructions, sequences were checked by DNA sequencing (GATC-Biotech, Konstanz, Germany) to ensure that no mutation was introduced during the PCR. *E. coli* DH5α (Life Technologies, Carlsbad, CA, US) was used for vector construction and amplification. The plasmids were digested with *Not*I and used for transformation of strain JMY1212 (MATA ura3–302 leu2–270-LEU2-zeta xpr2–322 Δlip2 Δlip7 Δlip8) (Bordes et al. 2007) by the lithium acetate method as described previously (Barth and Gaillardin 1996). The strains constructed in this study were deposited at the International Center for Microbial Resources, CIRM-Levures (http://www7.inra.fr/cirmlevures/).

Production of recombinant enzymes in Y. lipolytica

Erlenmeyer flasks (500 mL) containing medium $Y_1T_2O_3$ (50 mL total) made of yeast extract (10 g/L), bactotryptone (20 g/L), and oleic acid (30 g/L), buffered with phosphate buffer (100 mM, pH 6.8) were inoculated with cells pre-grown in YPD containing yeast extract (10 g/L), bactopeptone (10 g/L), and glucose (10 g/L) at an initial cell density of OD_{600}=0.2. Stock

solution containing oleic acid (200 g/L) and Tween40 (5 g/L) was subjected to sonication three times for 1 min each on ice for emulsification purposes. Cells were incubated at 28°C for 24 h until complete consumption of oleic acid. The cells were centrifuged at 10,000 rpm for 10 min, and supernatants were directly used in reactions.

SDS-PAGE analysis

Samples were loaded on 10% NuPAGE Tris-acetate sodium dodecyl sulphate-polyacrylamide electrophoresis gel (Invitrogen, Cergy Pontoise, France). After migration, the gel was subjected to silver staining.

Hydrolysis of p-nitrophenyl butyrate

Lipase activity in the culture supernatant was determined by monitoring the hydrolysis of *p*-nitrophenyl butyrate (*p*NPB) into butyric acid and *p*-nitrophenol (*p*-NP) (Quinn et al. 1982). The method was optimised using 2-methyl-butan-2-ol (2M2B) as solvent to solubilize *p*-NPB (Bordes et al. 2007). Lipase activity was measured in 96-well microplates with 20 µL of the supernatant containing properly diluted lipases, 175 µL of a 100 mM phosphate buffer pH 7.2 containing 100 mM NaCl and 5 µL of 40 mM pNPB in 2M2B. Activity was measured by monitoring absorbance at 405 nm at 25°C for 10 min using the VersaMax tunable microplate reader apparatus (Molecular Devices, Rennes, France). One unit of lipase activity was defined as the amount of enzyme releasing 1 µmol of pNP per min at 25°C and pH 7.2.

Resolution of (R, S) 2-bromo phenylacetic acid octyl esters

Hydrolysis was carried out in 2 mL Eppendorf tubes containing a biphasic medium composed of 0.75 mL dried decane containing the ester (50 mM) and 0.75 mL of the aqueous enzymatic solution. The mixture was shaken in a Vortex Genie 2 (Dutscher, Brumath, France). Reactions were performed at 25°C. After phase separation by centrifugation (50 µL diluted in 500 µL hexane), the progress of the reaction was monitored by taking samples at regular intervals.

HPLC analysis

The HPLC device was equipped with a chiral column: Chiralpack OJ (25 cm x 4.6 mm) (Daicel Chemical Industries Ltd, Japan) connected to a UV detector (at 254 nm). A flow rate of 1.0 mL/min and a 40°C column temperature were used. The mobile phase was composed of a mixture of *n*-hexane/isopropanol (80:20 v/v).

Determination of the enantiomeric excess (ee), conversion and enantioselectivity (E-value)

The enantioselectivity (*E*-value) was the ratio of the initial rate:

E = (viS / viR) for a S enantioselectivity

E = (viR / viS) for a R enantioselectivity.

ViS and viR are the initial rates of the S and R enantiomer consumption, respectively. The initial rate was determined by linear regression with a maximum conversion of substrate of less than 15%.

Results

A large family of lipases in the *Yarrowia* clade

Y. lipolytica is known to have undergone a lipase family expansion in a different way from most *Saccharomycotina* species. Indeed, 16 genes were recovered from the complete genome sequence (Dujon et al. 2004). We searched for homologues in the other sequenced species and found similar expansion in YAGA with 14 genes. Expansion was more limited in the other species, with 9 genes in YAYA (7 + 2 pseudogenes) and YAPH (8 + 1 pseudogene), 10 genes in YAAL (9 + 1 pseudogene), and only three genes in the out-group species CAHI. Thus, the set of LIP genes, the so called LIP family, is composed of 61 members of which 4 are pseudogenes (sequences are listed in Table S1). The two YAYA pseudogenes were shown to be full-length copies inactivated by a frameshift (YAYA0S2-34200g) or a stop in frame (YAYA0S4-04016g). The YAPH and YAAL pseudogenes corresponded to truncated copies (YAPH0S3-02454g and YAAL0S2-06656g). Full-length proteins ranged from 299 to 429 amino acids, the longest being YALI0D09064g (=YlLip11) with a putative N-terminal extension of 85 amino acids. Overall, the N-ter of the proteins was poorly conserved whereas the rest of the alignment was rather well conserved covering about two thirds of the proteins. All the proteins in this region harbored the conserved serine, aspartic acid, and histidine catalytic triad characteristic of the α/β hydrolase with the catalytic nucleophile serine located in the highly conserved pentapeptide GHS[LFM]G (Schrag and Cygler 1997). This finding confirms that all of these genes encode lipases. The most closely related lipases already published in other species are three lipases of *Candida deformans,* which also belongs to the *Yarrowia* clade (Bigey et al. 2003). The set of LIP genes, including those reported in *C. deformans*, but with the exception of *C. hispaniensis* LIP9, forms a monophyletic group sharing 32 to 97% identity. Outside this group, the proteins which share the highest level of identity correspond to putative lipases of *Wickerhamomyces ciferii* and *Kuraishia capsulata* with 31-33% identity and 51-52% similarity, respectively. Filamentous ascomycetes such as *Thermomyces lanuginosa*, *Penicillium, Blumeria,* or *Aspergillus* species as well as early-diverging fungus such as *Mucor miehei* or *Rhizopus niveus* contain LIP2 homologues which share about 25-30% identity and 40-50% similarity. A putative lipase was also identified in *Sacharomycetaceae*. This putative lipase in *Saccharomyces cerevisiae,* called Lih1 and encoded by YJR107W, shares 26% identity and 44% similarity with YlLip2, its closest relative in *Y. lipolytica*. Lih1 and its counterparts in other Hemiascomycetes may derive from a common LIP ancestor.

Reconstruction of an evolutionary scenario of the lipase family

To reconstruct the evolution of the lipase family, CAHI was used as an out-group and the common ancestor of the clade was placed after the divergence of CAHI, just before the branching of YAAL (Fig. 1a). To determine the phylogenetic relationships between the different lipases of the 6 species, an amino acid alignment was performed for the 59 full-length genes. The phylogeny based on this alignment made it possible to separate the proteins into 9 monophyletic groups with at least 4 members (Fig. 1b). In each group, we compared the phylogeny of the genes and the species, and we used synteny conservation to reconstruct the evolutionary scenario and thus to determine the number and the relative location of the ancestral lipases. This strategy showed that 15 duplications, 11 losses and 4 pseudogenizations may have occurred during the evolution of 10 ancestral genes (LIP2, LIP4, LIP5, LIP8, LIP9, LIP10, LIP13, LIP16, LIP18 and LIP18b). With on average more than 3 events per ancestral gene, the lipase family is thus highly dynamic in this clade with more duplications than losses. Assuming that CAHI has only three lipases, the family expansion may predate the divergence of the clade and is still underway in the most closely related species of YALI. Two duplications occurred before the divergence of YAYA/YAGA/YALI and three before the divergence of YAGA/YALI. All the other duplication events are species-specific (Fig. 1a). Finally in YALI, which has the highest number of lipases in the clade, LIP15 derived from the duplication of LIP4, LIP7 and LIP14 from LIP8, LIP11 and LIP12 from LIP10, LIP17 from LIP16 and LIP19 from LIP18, but only two duplications are species-specific (LIP12 and LIP19). The fact that 10 out of 11 losses are species-specific also contributed to the lipase family expansion. The 11th loss occurred before the divergence of YAPH and allowed the definition of a tenth ancestral lipase (LIP18b), absent from YALI and restricted to YAAL (Fig. 1b). We did not find any evidence for a putative PrePro domain in YaLIP18b, suggesting that the protein is not secreted. However, it would be interesting to investigate its intracellular activity and substrate specificity, if any.

The number of evolutionary events was not uniformly distributed among the phylogenetic groups of lipases. For instance, no event occurred in the LIP13 group and all the LIP13 genes are still conserved in synteny. On the contrary, the LIP2 gene has undergone multiple evolutionary events. Six independent duplications occurred, five in YAPH and one in YAAL. In both species, one non-ancestral copy is pseudogenized. However, in the 5 species which make up the *Yarrowia* clade, the ancestral copy of LIP2 is still conserved in synteny (Fig. 2).

Expression patterns are conserved only in the LIP2 family

YALI, YAGA and YAPH have at least 9 different genes, however only half of them (19 out of 39) were transcribed in the culture conditions of our study. No reads were obtained in the RNA-Seq data for 7 lipases of YALI and 7 of YAGA. In YAPH, none of the LIP2-like lipases except LIP2a which is the syntenic orthologue of YlLIP2 and thus the ancestral copy, were transcribed, neither the YpLIP18 lipase. Thus, only 9 lipases in YALI, 7 in YAGA and 3 in YAPH were considered to be transcriptionally active (Table S2). Their relative expression differed with the growth media. In the three strains, YlLIP2, YgLIP2 and YpLIP2 were the lipases most frequently expressed in oleic acid with more than 95% of the lipase transcripts in YALI and YAGA, and about 80% in YAPH (Fig. 3a). Surprisingly, in YALI the major lipase in glucose was YlLIP8 whereas it was still YgLIP2 in YAGA. In YAPH, the expression of YpLIP13 prevailed in both glucose and tributyrin. In these media, there was no predominant lipase in YALI. Among the 19 expressed lipases, at least 11 lipases appeared to be regulated, i.e. they showed statistically significant difference in levels of expression in at least one media comparison (see the adjusted P-value in Table S2 and Fig. 3b). The three expressed LIP2 lipases showed the same pattern of expression regulation. These lipases were expressed in the three media and were mainly induced on oleic acid and slightly less on glucose than on tributyrin. Conservation of the pattern of regulation between orthologous lipases is not a general rule. Indeed, YlLIP4 was not expressed in any condition; YpLIP4 was mainly induced on tributyrin and YgLIP4 on oleic acid. Similarly, YpLIP13 was expressed in all three conditions; YlLIP13 was induced on glucose and YgLIP13 on oleic acid. In YALI, four additional lipases were shown to be regulated by growth conditions (LIP8, LIP9, LIP10 and LIP17); none of these lipases was found in YAPH but were present in YAGA. LIP10 and LIP17 were induced by tributyrin in YALI and not expressed in YAGA. LIP9 was induced by tributyrin in YALI and was always expressed in YAGA.

A promising strong promoter inducible by glucose

The most remarkable regulation was that of YlLIP8 which was highly induced by glucose with fold change values of 360 and 562 for glucose versus oleic acid and tributyrin, respectively. A comparison of the level of expression between YlLIP8 and all other protein-coding genes of *Y. lipolytica* showed that YlLIP8 is one of the 25 most frequently expressed genes on glucose (read counts per kb in both glucose replicates). This gene is almost never expressed on oleic acid or tributyrin. Its expression pattern in our conditions resembled that of the XPR2 promoter (pXPR2) from YALI0F31889g (Madzak et al. 1999). pXPR2 is a strong inducible promoter with a complex induction (Ogrydziak et al. 1977), which has encouraged

people to use hybrid chimera promoters based on the XPR2 upstream activation sequences (Blazeck et al. 2011; Madzak et al. 2000). Assuming that pLIP8 has the same expression and specificity as pXPR2, it may be an interesting alternative promoter with less complex requirements for induction, and deserves further investigation.

Structural characteristics of the LIP2 subfamily

The LIP2 gene is the only lipase gene present in all species of the clade and in CAHI. This gene has undergone the highest number of duplications and the ancestral copy is the most frequently expressed lipase in oleic acid in YALI, YAGA and YAPH. All these findings mean this subfamily, the so- called LIP2 family, would be an interesting group of genes in which to investigate their respective activity and enantioselectivity in relation with their sequence features.

The sequence alignment of the 10 newly identified Lip2 lipases with YlLIP2 and the Lip2 homologue previously identified in *Y. deformans* (Bigey et al. 2003) is shown in Fig. 4a. Compared to YlLip2, the YgLip2, YdLip2 and YyLip2a lipases present the highest level of conservation with 87.7 to 92.5% pairwise identity and 92.8 to 97.3% similarity, which is congruent with the species phylogeny. In contrast, the higher conservation of YlLip2 with YaLip2 (87.7% identity and 93.1% similarity) rather than with YpLip2a, which is the ancestral copy in YAPH, is surprising (Fig. 4b). As both genes are still conserved in synteny (Fig. 2), the high level of sequence conservation of YaLip2 may be due to a horizontal gene transfer followed by a conversion event. A possible alternative explanation is that YpLip2a diverged more rapidly due to the presence of 5 duplicated copies in YAPH, even though, among the YAPH LIP2 lipases, YpLip2a, which is the ancestral copy, remains the most conserved. To test this alternative hypothesis, we quantified the selective constraints on the LIP2 genes by estimating the ratio of per site non-synonymous (dN) to synonymous (dS) substitution rates (dN/dS). As shown in Fig. 5, all LIP2 genes were shown to be under purifying selection, i.e. the dN/dS ratio was less than 1. The pressure acting on non-ancestral LIP2 genes was much lower than on ancestral LIP2 pairs, which might be due to a relaxation of the purifying selection on duplicated copies. Although multiple copies exist in YAPH, the ancestral YpLIP2a gene was seen to be under the same purifying selection pressure as ancestral copies in the other genomes where no LIP2 duplications occurred. This finding supports the first hypothesis of a horizontal transfer at the locus in YAAL.

Also noteworthy is the fact that YpLIP2b and YpLip2c are almost identical with only 11 amino acid differences, suggesting that one is a duplicated copy of the other. In contrast, the two LIP2 copies in CAHI shared only 55.9 % identity and 69.3% similarity, probably

resulting from a long evolution. Sequencing the LIP2 locus of additional species of the *Yarrowia* clade, especially those which have been described very recently, will provide clues to decipher the strange phylogeny of the LIP2 gene (Chang et al. 2013; Nagy et al. 2014; Nagy et al. 2013).

The secondary structure pattern appeared to be highly conserved among all 12 enzymes, with a perfect alignment of α-helices and β-strands. The classical signature GHSLG of lipases from the filamentous fungi superfamily, which includes the catalytic serine, was seen to be almost conserved. The only exception was YpLip2d from *Y. phangngensis* in which the leucine next to the catalytic serine was replaced by a phenylalanine. The two other amino acids of the catalytic triad, namely Asp230 and His289 in YlLip2, were perfectly aligned located after strand β7 and β9, respectively. The two residues forming the oxyanion hole, which stabilize the tetrahedral intermediate formed during the reaction mechanism, were aligned and identical. The first residue belongs to the "GX" type lipase according to the classification proposed by Pleiss *et al.*, which bears specificities for medium and long chain fatty acids, X being a threonine in all these lipases (Pleiss et al. 2000). The second residue, located next to the catalytic serine, was a leucine, except for the previously mentioned YpLip2d with a phenylalanine.

Disulphide bonds play an essential structural role in proteins by stabilizing their tertiary structure. YlLip2 is stabilised by four disulphide bonds (Cys30-Cys299, Cys43-Cys47, Cys120-Cys123, Cys265-Cys273), and contains one free cysteine (Cys244) (Bordes et al. 2010). All the lipases from the LIP2 family appeared to be stabilised by the same four disulphide bonds. The extra cysteine, Cys244 reported in YlLip2, was present in all ancestral LIP2 lipases of the clade, i.e. in YpLip2a, YgLip2, YyLip2 and YaLip2, but not in YhLip2a,b. This extra cysteine has been shown to be detrimental to enzyme thermostability in YlLip2, leading to intramolecular disulphide bonds during denaturation and protein aggregation (Bordes et al. 2011).

One specificity of lipases is the fact that catalytic activity is strongly dependent on the structural rearrangement of a mobile sub-domain, called the lid. The lid was seen to block the active site in the closed form of the enzyme and to move away in presence of a hydrophobic interface, and had an open conformation when the active site became accessible to the substrate. Generally, the sequence of the lid (from L91 to I100 in YlLip2) of the different lipases was similar. YhLip2a and YhLip2b had the most dissimilar lid sequence.

The substrate binding site of YlLip2 appeared as a hydrophobic crevice located at the protein surface, with a catalytic triad exposed to the solvent, like in all mucorales lipases. The hydrophobic crevice consisted of T88, V94, I98, I100, F129, L163, P190, V232, V235, P236,

Y241 in YlLip2. The scissile fatty acid of a triglyceride is assumed to bind to this hydrophobic crevice. The sn-2 substituent is thought to bind to the hydrophobic dent formed by I231, V283, V285 and L290. Table 2 lists the amino acids located at these positions for the 12 Lip2 lipases. Amino acids 129, 190, 232, 236 and 241 were conserved in the 15 positions. The first amino acid of the oxyanion hole in position 88 was a threonine, except in YhLip2b, where it was a serine like in other mucorales lipases such as *Mucor miehei* and *Thermomyces lanuginosa* (Bordes et al. 2011). The second amino acid of the oxyanion hole was a leucine except for YpLip2d where it was a phenylalanine. Amino acids in positions 98 and 100, located in the lid, and in 231, 235, 285 and 290, were all hydrophobic amino acids (V, L, I, A or F). Positions 94 and 283 presented the highest variability. In position 94, hydrophobic amino acids are generally encountered, with a tryptophan in YhLip2b, but a threonine was found in YpLip2e. The amino acid in position 283 was a hydrophobic residue except for YhLip2a and YhLip2b, where it was a lysine, a positively charged amino acid.

Finally, YlLip2, as an excreted protein, was expressed in the form of a PrePro enzyme precursor (Fig. 4a, Fig. S1). The pre region is composed of 22 amino acids with four X-Ala or X-Pro dipeptides, substrates of a diamino peptidase which cleaves after Ile22 (Pignede et al. 2000). It is followed by a short pro region of 12 aa finished by a Lys-Arg dipeptide, substrate of the *KEX2*-like endopeptidase encoded by the *XPR6* gene in *Y. lipolytica* (Pignede et al. 2000). YpLip2d had no clear PrePro region. YhLip2a and b PrePro regions diverged from the YlLip2 PrePro region. As shown on the alignment of the 9 most closely related PrePro regions to that of YlLip2, two X-Arg dipeptides were present at the end of most of these PrePro regions, except in YaLip2, YdLip2, YgLip2 and YpLip2a, where only the Arg-Arg dipeptide was observed (Fig. S1). Two X-Pro were also present in the PrePro of these 9 lipases except in YyLip2 where the second dipeptide, Ser-Pro, was replaced by a Thr-Ala.

Expression of LIP2 genes in *Y. lipolytica*

All LIP2 genes except YdLIP2 were cloned in *Y. lipolytica* strain JMY1212 (Bordes et al. 2007) under the control of the constitutive TEF promoter from *Y. lipolytica*. Two strategies were used: either the entire genes were cloned in the presence of their own PrePro sequences or the part of the genes coding for the mature proteins were cloned behind the PrePro region of YlLip2. Hydrolytic activities of *p*Nitro Phenyl Butyrate (*p*NPB) were measured and are presented in table 3.

Seven lipases, YlLip2, YaLip2, YpLip2a, YpLip2b, YpLip2c, YgLip2 and YyLip2, presented significant activities whatever the strategy used, suggesting that the origin of the PrePro region has no clear influence on the result. The other lipases, YpLip2d, YpLip2e, YhLip2a

and YhLip2b, showed no or very low activity when expressed in their own PrePro region. Except YpLip2e, all had a different N-terminal region from the PrePro of other LIP2 members, but even when the latter were replaced by the PrePro region of YlLip2, no or very little activity was detected. This suggests that either the protein was not correctly secreted in the medium, or degraded rapidly, or that the secreted protein no longer showed lipase activity. The presence of a secreted protein was checked on a protein gel. This experiment showed that all 7 active enzymes presented similar expression, whereas in the case of no or low activity, no protein expression was detected whatever the PrePro system used (Fig. S2). This absence of protein is surprising and may reflect a problem in the coding sequence, as in each case two different proteins were produced, one with YlLip2 PrePro and one with the native PrePro. For YpLip2d, we additionally constructed the F165L variant with YlLip2 PrePro. Indeed, YpLip2d presented extra specificity, as it had a bulky atom in the oxyanion hole, a phenylalanine, instead of a leucine. This F165L variant was constructed but the activity was not restored, suggesting that the absence of activity is not solely due to this amino acid substitution.

Interestingly, in YAPH, three Lip2 lipases showed significant enzymatic activity under the TEF promoter, but only YpLip2a was expressed in the wild type strain under its own promoter in the conditions used for the RNA-Seq analysis. Indeed, the genes coding for YpLip2b and YpLip2c were not transcribed in the wild type strain but were well expressed when controlled by the TEF promoter.

Enantioselectivity of the LIP2 subfamily

The enantioselectivity of 11 lipases from the LIP2 family was tested during the resolution by hydrolysis of 2-bromo-arylacetic acid esters (Table 4). YpLip2a was clearly the most efficient lipase from both kinetic and selectivity points of view. It displayed 2–fold enhanced activity toward the *S*-enantiomer whereas its activity toward the *R*-enantiomer was about 5 times lower than that obtained with the YlLip2. As a consequence, enantioselectivity increased remarkably, by almost one order of magnitude, from an *E*-value of 3 to 30, compared to other lipases.

Position 97 and mainly position 232 were identified as crucial for distinguishing between enantiomers. However, these two positions are perfectly conserved with a valine and an aspartic acid in positions 232 and 97, respectively. This means that the higher observed enantioselectivity is due to a more subtle change in the 3D structure of this lipase.

In a previous work, screening of a library of saturation of position 232 led to the identification of the V232S variant, with an E-value tremendously increased compared to the parental

enzyme (E-value=230) (Bordes et al. 2009). Valine at position 232 was changed by a serine inYpLip2a. After 12 hours of reaction, the preferred *S*-enantiomer was completely consumed whereas the concentration of the *R*-enantiomer remained unchanged (Fig. 6). In addition to the gain in enantioselectivity, a 2.5 increase in velocity was observed (Table 4). With this variant, both enantiomers can be recovered with 100% purity and 100% yield. Finally, based on a previous work on YlLip2 (Cambon et al. 2010), a double-substituted variant was constructed in YpLip2a by site-directed mutagenesis, the variant D97A-V232F. This latter presented reverse enantioselectivity with a total preference for the *R*-enantiomer (Table 4).

Discussion

The objective of this work was to test an original idea to maximize the success of the design of an optimal enzyme. The chances of success are considerably increased when the selected enzyme naturally performs well in terms of activity and selectivity.

Lip2p lipase from *Yarrowia lipolytica* (Yllip2) was shown to be an efficient catalyst for the resolution of 2-bromo-arylacetic acid esters, an important class of chemical intermediates in the pharmaceutical industry (Guieysse et al. 2004). However, although its activity is satisfactory, its enantioselectivity needs to be improved. Our strategy, before starting enzyme engineering, was to look for a better enzyme which could be produced by a member of the *Yarrowia* clade. The strength of our analysis was to successfully reconstruct the scenario for the evolution of the LIP family which gave rise to 16 lipases in YALI. The number of evolutionary events was not uniformly distributed among the different groups of lipases. The LIP2 family has undergone the highest number of evolutionary events, leading to identifying 11 lipases homologous to YlLip2 among the 59 putative lipases found in this study. In *Y. phangngensis*, five duplications of *LIP2* took place, which constitutes a reservoir for the detection of promising enzymes. One surprising finding of the high-throughput transcriptome analysis was the fact that none of the duplicated copies of *LIP2* was expressed in the conditions tested. From a fundamental point of view, this result deserves to be investigated at a larger scale, i.e. in all expanded protein families of these genomes.

Ten out of the 11 Lip2 lipases were cloned and expressed in *Y. lipolytica*, but only six were active using the classical *pNPB* hydrolysis test, whatever the target PrePro-peptide used for secretion. Three of the active lipases belong to *Y. phangngensis,* whereas only YpLip2a was found to be expressed in the wild type strain under the conditions tested. The fact that the enzymatic activity of the two non ancestral YpLip2b and YpLip2c has been conserved across evolution implies a selective pressure acting on these genes, otherwise this activity would have been lost, as in the case of YpLIP2d and YpLIP2e. The evolutionary scenario revealed that YpLIP2b and YpLIP2c derived one from the other due to a recent duplication. The alignment of the flanking regions of these two genes showed that the duplication event included both the promoter and terminator. From an evolutionary point of view, it is interesting to note that the duplicated region exceeds the transcribed part of the gene, suggesting that the mechanism responsible for the duplication did not involve an RNA intermediate, but rather a segmental DNA duplication (Fig. S3). From this evolutionary analysis, we can deduce that YpLIP2b and YpLIP2c probably have the same expression regulation but the conditions of induction remain to be determined.

The Lip2 lipase from *Y. lipolytica* has been reported to be an efficient stereoselective enzyme for the resolution of 2-halogeno-arylacetic acid esters. In our study, the screening for enantioselectivity and activity led to the identification of the *Yarrowia phangngensis* lipase, Yplip2a, which has a tremendously increased E-value compared to the first identified YlLip2 for the resolution of 2-bromo-phenylacetic acid ethyl ester (9-fold, E-value 30). In addition to the gain in enantioselectivity, a 2-fold increase in velocity was observed. Finally, the best candidate was subjected to enzyme engineering by site-directed mutagenesis targeted to the active site. The mono-substituted variant V232S showed significantly enhanced selectivity (E-value > 200) compared to wild-type enzyme. The improved E-value was the consequence of an increase in the reaction rate of the fast-reacting enantiomer, leading to a 2.5 increase in velocity. This variant performed better than the corresponding variant of YlLip2 in terms of enantioselectivity due to complete non-recognition of the non-preferred enantiomer (Bordes et al. 2009). A second double-substituted variant D97A-V232F of YpLip2a presented reversed enantioselectivity, with a total preference for the *R*-enantiomer, as already observed with YlLip2 (Cambon et al. 2010). The two YpLip2a variants are now compatible with industrial applications in the pharmaceutical industry.

This study is a clear example of the power of using natural biodiversity exploited by comparative genomics for enzyme property optimization to fulfil industrial requirements. Facilitated by the advances in high-throughput sequencing technologies, this strategy should become widespread in the next few years.

Acknowledgements

We thank Claude Gaillardin for stimulating discussions and Serge Casaregola from the CIRM-Levures for the storage of our LIP2-expressing strains. This work was performed, in partnership with the SAS PIVERT, within the frame of the French Institute for the Energy Transition (Institut pour la Transition Energétique - ITE) P.I.V.E.R.T. (www.institut-pivert.com) selected as an Investment for the Future ("Investissements d'Avenir"). This work was supported as part of the Investments for the Future, by the French Government under the reference ANR-001-01 and by INRA through the AIP-Bioressources YALIP project. Muchalin Meunchan acknowledges the Thailand Research Fund (TRF) together with Khon Kaen University for financial support through Royal Golden Jubilee PhD program (grant number PHD/0282/2549) and French Embassy. Stéphanie Michely has a PhD fellowship funded by the French Minister of Research via the ABIES doctoral school.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403-10 doi:10.1016/S0022-2836(05)80360-2

Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biology 11(10):R106

Barth G, Gaillardin C (1996) *Yarrowia lipolytica*. In: Wolf WK (ed) Non-Conventional Yeasts in Biotechnology. vol 1. Springer-Verlag, Berlin, pp 313-388

Benjamin S, Pandey A (1998) *Candida rugosa* lipases: molecular biology and versatility in biotechnology. Yeast 14(12):1069-87

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) 57(1):289-300 doi:doi: 10.2307/2346101

Bigey F, Tuery K, Bougard D, Nicaud J-M, Moulin G (2003) Identification of a triacylglycerol lipase gene family in *Candida deformans*: molecular cloning and functional expression. Yeast 20(3):233-248 doi:10.1002/yea.958

Blazeck J, Liu L, Redden H, Alper H (2011) Tuning gene expression in *Yarrowia lipolytica* by a hybrid promoter approach. Appl Environ Microbiol 77(22):7905-14 doi:10.1128/AEM.05763-11

Bordes F, Barbe S, Escalier P, Mourey L, Andre I, Marty A, Tranier S (2010) Exploring the conformational states and rearrangements of *Yarrowia lipolytica* Lipase. Biophys J 99(7):2225-34 doi:10.1016/j.bpj.2010.07.040

Bordes F, Cambon E, Dossat-Letisse V, Andre I, Croux C, Nicaud JM, Marty A (2009) Improvement of *Yarrowia lipolytica* lipase enantioselectivity by using mutagenesis targeted to the substrate binding site. Chembiochem 10(10):1705-13 doi:10.1002/cbic.200900215

Bordes F, Fudalej F, Dossat V, Nicaud JM, Marty A (2007) A new recombinant protein expression system for high-throughput screening in the yeast *Yarrowia lipolytica*. J Microbiol Methods 70(3):493-502 doi:10.1016/j.mimet.2007.06.008

Bordes F, Tarquis L, Nicaud JM, Marty A (2011) Isolation of a thermostable variant of Lip2 lipase from *Yarrowia lipolytica* by directed evolution and deeper insight into the denaturation mechanisms involved. J Biotechnol 156(2):117-24 doi:10.1016/j.jbiotec.2011.06.035

Buchan DW, Minneci F, Nugent TC, Bryson K, Jones DT (2013) Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Res 41(Web Server issue):W349-57 doi:10.1093/nar/gkt381

Cambon E, Piamtongkam R, Bordes F, Duquesne S, Andre I, Marty A (2010) Rationally engineered double substituted variants of *Yarrowia lipolytica* lipase with enhanced activity coupled with highly inverted enantioselectivity towards 2-bromo phenyl acetic acid esters. Biotechnol Bioeng 106(6):852-9 doi:10.1002/bit.22770

Cardenas F, de Castro MS, Sanchez-Montero JM, Sinisterra JV, Valmaseda M, Elson SW, Alvarez E (2001) Novel microbial lipases: catalytic activity in reactions in organic media. Enzyme Microb Technol 28(2-3):145-154

Casas-Godoy L, Meunchan M, Cot M, Duquesne S, Bordes F, Marty A (2014) *Yarrowia lipolytica* lipase Lip2: An efficient enzyme for the production of concentrates of docosahexaenoic acid ethyl ester. J Biotechnol 180C:30-36 doi:10.1016/j.jbiotec.2014.03.018

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17(4):540-52

Chang CF, Chen CC, Lee CF, Liu SM (2013) Identifying and characterizing *Yarrowia keelungensis sp. nov*., an oil-degrading yeast isolated from the sea surface microlayer. Antonie Van Leeuwenhoek 104(6):1117-23 doi:10.1007/s10482-013-0033-z

Corpet F (1988) Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res 16(22):10881-90

de Almeida AF, Tauk-Tornisielo SM, Carmona EC (2013) Acid lipase from *Candida viswanathii*: production, biochemical properties, and potential application. Biomed Res Int 2013:435818 doi:10.1155/2013/435818

Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, de Montigny J, Marck C, Neuveglise C, Talla E, Goffard N, Frangeul L, Aigle M, Anthouard V, Babour A, Barbe V, Barnay S, Blanchin S, Beckerich J-M, Beyne E, Bleykasten C, Boisrame A, Boyer J, Cattolico L, Confanioleri F, de Daruvar A, Despons L, Fabre E, Fairhead C, Ferry-Dumazet H, Groppi A, Hantraye F, Hennequin C, Jauniaux N, Joyet P, Kachouri R, Kerrest A, Koszul R, Lemaire M, Lesur I, Ma L, Muller H, Nicaud J-M, Nikolski M, Oztas S, Ozier-Kalogeropoulos O, Pellenz S, Potier S, Richard G-F, Straub M-L, Suleau A, Swennen D, Tekaia F, Wesolowski-Louvel M, Westhof E, Wirth B, Zeniou-Meyer M, Zivanovic I, Bolotin-Fukuhara M, Thierry A, Bouchier C, Caudron B, Scarpelli C, Gaillardin C, Weissenbach J, Wincker P, Souciet J-L (2004) Genome evolution in yeasts. Nature 430(6995):35-44

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113 doi:10.1186/1471-2105-5-113

Fickers P, Fudalej F, Dall MTL, Casaregola S, Gaillardin C, Thonart P, Nicaud JM (2005) Identification and characterisation of LIP7 and LIP8 genes encoding two extracellular triacylglycerol lipases in the yeast *Yarrowia lipolytica*. Fungal Genetics and Biology 42(3):264-274

Fickers P, Marty A, Nicaud JM (2011) The lipases from *Yarrowia lipolytica*: genetics, production, regulation, biochemical characterization and biotechnological applications. Biotechnol Adv 29(6):632-44 doi:10.1016/j.biotechadv.2011.04.005

Gopinath SC, Anbu P, Lakshmipriya T, Hilda A (2013) Strategies to characterize fungal lipases for applications in medicine and dairy industry. Biomed Res Int 2013:154549 doi:10.1155/2013/154549

Guieysse B, Salagnad C, Monsan P, Remaud-Simeon M (2003) Lipase-catalyzed enantioselective transesterification toward esters of 2-bromo-tolylacetic acids. Tetrahedron: Asymmetry 14(3):317-323

Guieysse B, Sandoval G, Faure L, Nicaud JM, Monsan P, Marty A (2004) New efficient lipase from *Yarrowia lipolytica* for the resolution of 2-bromo-arylacetic acid esters. Tetrahedron: Asymmetry 15(22):3539-3543

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52(5):696-704

Hube B, Stehr F, Bossenz M, Mazur A, Kretschmar M, Schafer W (2000) Secreted lipases of *Candida albicans*: cloning, characterisation and expression analysis of a new gene family with at least ten members. Arch Microbiol 174(5):362-74

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30(14):3059-66

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14(4):R36 doi:10.1186/gb-2013-14-4-r36

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol 5(2):R12 doi:10.1186/gb-2004-5-2-r12

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078-9 doi:10.1093/bioinformatics/btp352

Madzak C, Blanchin-Roland S, Cordero Otero RR, Gaillardin C (1999) Functional analysis of upstream regulating regions from the *Yarrowia lipolytica* XPR2 promoter. Microbiology 145 (Pt 1):75-87

Madzak C, Treton B, Blanchin-Roland S (2000) Strong hybrid promoters and integrative expression/secretion vectors for quasi-constitutive expression of heterologous proteins in the yeast *Yarrowia lipolytica*. J Mol Microbiol Biotechnol 2(2):207-16

Masse L, Kennedy KJ, Chou S (2001) Testing of alkaline and enzymatic hydrolysis pretreatments for fat particles in slaughterhouse wastewater. Bioresour Technol 77(2):145-55

Nagy E, Dlauchy D, Medeiros AO, Peter G, Rosa CA (2014) *Yarrowia porcina sp. nov.* and *Yarrowia bubula f.a. sp. nov.*, two yeast species from meat and river sediment. Antonie Van Leeuwenhoek doi:10.1007/s10482-014-0125-4

Nagy E, Niss M, Dlauchy D, Arneborg N, Nielsen DS, Peter G (2013) *Yarrowia divulgata f.a., sp. nov.*, a yeast species from animal-related and marine sources. Int J Syst Evol Microbiol 63(Pt 12):4818-23 doi:10.1099/ijs.0.057208-0

Ogrydziak DM, Demain AL, Tannenbaum SR (1977) Regulation of extracellular protease production in *Candida lipolytica*. Biochim Biophys Acta 497(2):525-38

Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, Harel M, Remington SJ, Silman I, Schrag J, Sussman JL, Verschueren KHG, Goldman A (1992) The alpha/beta hydrolase fold. Protein Eng 5(3):197-211

Piamtongkam R, Duquesne S, Bordes F, Barbe S, Andre I, Marty A, Chulalaksananukul W (2011) Enantioselectivity of *Candida rugosa* lipases (Lip1, Lip3, and Lip4) towards 2-bromo phenylacetic acid octyl esters controlled by a single amino acid. Biotechnol Bioeng 108(8):1749-56 doi:10.1002/bit.23124

Pignede G, Wang H, Fudalej F, Gaillardin C, Seman M, Nicaud JM (2000) Characterization of an extracellular lipase encoded by LIP2 in *Yarrowia lipolytica*. J Bacteriol 182(10):2802-10

Pleiss J, Fischer M, Peiker M, Thiele C, Schmid RD (2000) Lipase engineering database: Understanding and exploiting sequence–structure–function relationships. Journal of Molecular Catalysis 10(5):491-508

Quinn DM, Shirai K, Jackson RL, Harmony JA (1982) Lipoprotein lipase catalyzed hydrolysis of water-soluble p-nitrophenyl esters. Inhibition by apolipoprotein C-II. Biochemistry 21(26):6872-9

R Development Core Team (2011) R: A language and environment for statistical computing. In: Computing RFfS (ed). Vienna, Austria.

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16(6):276-7

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139-40 doi:10.1093/bioinformatics/btp616

Schrag JD, Cygler M (1997) Lipases and alpha/beta hydrolase fold. Methods Enzymol 284:85-107

Tan T, Lu J, Nie K, Deng L, Wang F (2010) Biodiesel production with immobilized lipase: A review. Biotechnol Adv 28(5):628-34 doi:10.1016/j.biotechadv.2010.05.012

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24(8):1586-91 doi:10.1093/molbev/msm088

Figure legend

Figure 1: Phylogeny and evolutionary scenario of the lipase genes in the *Yarrowia* clade.
A. Evolutionary events are placed on the species tree constructed from the concatenation of 912 proteins (398959 residues). Duplications, losses and pseudogenizations are depicted as vertical green, red and blue lines, respectively. The number of ancestral genes at each ancestral node is given within yellow circles.
B. Phylogeny of the lipases from the *Yarrowia* clade and *C. hispaniensis*. The tree is based on the alignment of 59 lipases, further cleaned with Gblocks (133 amino acid residues). The two full-length pseudogenes YyLIP9p and YyLIP5p are included in the tree, whereas the truncated pseudogenes are not. Groups of lipases derived from the 10 ancestral copies are depicted as teardrops of different colours. The tenth ancestral lipase, called LIP18b, is present only in YAAL (YaLIP18b in bold, included in the LIP18 group). Bootstrap values are represented at the most ancestral nodes.

Figure 2: Synteny conservation between LIP2 genes. Orthologous genes are represented by arrows of the same colour. The name of the *Y. lipolytica* ortholog is abbreviated above the first line of genes or under the non-orthologous genes that are depicted in grey. The inversion event that occurred between YAYA S6 and YAAL S10 is represented with two pink inverted triangles. Chromosome or scaffold abbreviations on the left correspond to *Y. lipolytica* chromosome A (YALI A), *C. galli* chromosome E (YAGA E), *Y. yakushimensis* scaffold 6 (YAYA S6), *C. alimentaria* scaffold 10 (YAAL S10), *C. phangngensis* scaffold 6 (YAPH S6) and *C. hispaniensis* chromosome B (CAHI B).

Figure 3: Relative expression of lipases according to growth media in YALI, YAGA and YAPH.
A. The number of reads has been normalized by the length of the CDS for each gene and expressed as RPK (reads per kb). LIP name correspond to orthologous genes. In the case of *C. phangngensis*, the LIP2 corresponds to LIP2a which is the syntenic ortholog.
B. The fold change (FC) between each couple of conditions is provided for each expressed lipase of YALI, YAGA and YAPH with a significant adjusted P-value in at least one couple of conditions. Stars indicate fold change with a significant adjusted P-value (cut-off of adjP-value is 1.e-3 for YALI and 5.e-3 for YAGA and YAPH). GL: glucose, AO: oleic acid, TB: tributyrin.

Figure 4: Amino acids sequence comparison of lipases from the LIP2 family.
A. Amino acids sequence alignment. Residues forming part of α helices and β strands are coloured in magenta and cyan, respectively. The three catalytic residues are coloured in red, the two catalytic residues of the oxyanion hole in dark blue, the cysteines are yellow-coloured and the lid is in grey above the alignment. The α/β hydrolase fold of YlLip2 is from Figure 3 of Bordes et al. (Bordes et al. 2010). β-sheet are numbered according to the current nomenclature (Ollis et al. 1992). The YdLip2 corresponds to YdLIP1 (AJ428393.1; Bigey et al. 2003).
B. Pairwise sequence comparison. Similarity and identity are calculated along the whole pariwise alignment and provided as percentage. Colours are according to the level of conservation: below 50% (blue), 50-60% (cyan), 60-70% (green), 70-80% (yellow), 80-90% (orange) and 90-100% (red).

Figure 5: Box-plot of the dN/dS ratio estimated for the LIP2 genes based on pairwise alignments. Different sets of genes were compared: pairs of ancestral LIP2 genes (A), pairs of YpLIP2 genes (B) and ancestral versus non-ancestral LIP2 genes (C). Red dots represent the dN/dS values involvingYpLIP2a.

Figure 6: Hydrolysis kinetics of R,S-2-bromo-phenylacetic acid ethyl ester in a biphasic medium (water/decane v/v) at 25 °C. Activities of YlLip2 (green triangle) and YpLip2a (red diamond), over-expressed in *Y. lipolytica* YJM1212, towards S- (filled symbol) and R- (open symbol) enantiomers.
A. Activities of the wild-type YlLip2 and YpLip2a.
B. Activities of V232S variant of YlLip2 and YpLip2a.

Table 1 Strains used in this study

Strains used for evolutionary analysis		
Species	**Strain number**	**Reference**
Yarrowia lipolytica	CBS 7504=W29	(Barth and Gaillardin 1996)
Candida galli	CBS 9722	(Péter, et al. 2004)
Yarrowia yakushimensis	CBS 10253	(Groenewald and Smith 2013)
Candida phangngensis	CBS 10407	(Limtong, et al. 2008)
Candida alimentaria	CBS 10151	(Knutsen, et al. 2007)
Candida hispaniensis	CBS 9996	(Kurtzman 2005)

Strains used for lipase expression in *Y. lipolytica*			
Host species (lipase expressed)	**CLIB n°**	**Genotype**	**Reference**
Y. lipolytica JMY1212	CLIB 1677	*MATA ura3-302 leu2-270-LEU2-zeta xpr2-322 Δlip2 Δlip7 Δlip8* Leu$^+$, Ura$^-$	(Bordes, et al. 2007)
Y. lipolytica (*YlLIP2*)	CLIB 1678	JMY1212 + *pTEF-YlLIP2-URA3ex*	(Bordes, et al. 2011)
Y. lipolytica (*YlLIP2*-V232S)	CLIB 1679	JMY1212 + *pTEF-YlLIP2-URA3ex*	(Bordes, et al. 2009)
Y. lipolytica (*YlLIP2*-D97A-V232F)	CLIB 1680	JMY1212 + *pTEF-YlLIP2-URA3ex*	(Cambon, et al. 2010)
Y. lipolytica (*YpLIP2a*)	CLIB 1681	JMY1212 + *pTEF-YpLIP2a-URA3ex*	This study
	CLIB 1682	JMY1212 + *pTEF-YlPrePro-YpLIP2a-URA3ex**	This study
Y. lipolytica (*YpLIP2a*-V232S)	CLIB 1683	JMY1212 + *pTEF-YpLIP2a-URA3ex*	This study
Y. lipolytica (*YpLIP2a*-D97A-V232S)	CLIB 1684	JMY1212 + *pTEF-YpLIP2-URA3ex*	This study
Y. lipolytica (*YpLIP2b*)	CLIB 1685	JMY1212 + *pTEF-YpLIP2b-URA3ex*	This study
	CLIB 1686	JMY1212 + *pTEF-YlPrePro-YpLIP2b-URA3ex**	This study
Y. lipolytica (*YpLIP2c*)	CLIB 1687	JMY1212 + *pTEF-YpLIP2c-URA3ex*	This study
	CLIB 1688	JMY1212 + *pTEF-YlPrePro-YpLIP2c-URA3ex**	This study
Y. lipolytica (*YpLIP2d*)	CLIB 1689	JMY1212 + *pTEF-YpLIP2d-URA3ex*	This study
	CLIB 1690	JMY1212 + *pTEF-YlPrePro-YpLIP2d-URA3ex**	This study
Y. lipolytica (*YpLIP2e*)	CLIB 1691	JMY1212 + *pTEF-YpLIP2e-URA3ex*	This study
	CLIB 1692	JMY1212 + *pTEF-YlPrePro-YpLIP2e-URA3ex**	This study
Y. lipolytica (*YgLIP2*)	CLIB 1693	JMY1212 + *pTEF-YgLIP2-URA3ex*	This study
	CLIB 1694	JMY1212 + *pTEF-YlPrePro-YgLIP2-URA3ex**	This study
Y. lipolytica (*YyLIP2*)	CLIB 1695	JMY1212 + *pTEF-YyLIP2-URA3ex*	This study
	CLIB 1696	JMY1212 + *pTEF-YlPrePro-YyLIP2-URA3ex**	This study
Y. lipolytica (*YaLIP2*)	CLIB 1697	JMY1212 + *pTEF-YaLIP2-URA3ex*	This study
	CLIB 1698	JMY1212 + *pTEF-YlPrePro-YaLIP2-URA3ex**	This study
Y. lipolytica (*YhLIP2a*)	CLIB 1699	JMY1212 + *pTEF-YhLIP2a-URA3ex*	This study
	CLIB 1700	JMY1212 + *pTEF-YlPrePro-YhLIP2a-URA3ex**	This study
Y. lipolytica (*YhLIP2b*)	CLIB 1701	JMY1212 + *pTEF-YyLIP2b-URA3ex*	This study
	CLIB 1702	JMY1212 + *pTEF-YlPrePro-YhLIP2b-URA3ex**	This study

CBS: Centraalbureau voor Schimmelcultures, Utrecht, The Netherlands
CLIB numbers refer to the International Center for Microbial Resources, CIRM-Levures (http://www7.inra.fr/cirmlevures/)
*YlPrePro-Yx*LIP2* corresponds to the fusion of the *Y. lipolytica LIP2* PrePro region with the region of the expressed lipase corresponding to the mature coding region.

Table 2: Amino acids involved in the hydrophobic crevice and hydrophobic dent

	YlLip2	YpLip2a	YpLip2b	YpLIP2c	YpLip2d	YpLip2e	YgLip2	YdLip2	YyLip2	YaLip2	YhLip2a	YhLip2b
88	T	T	T	T	T	T	T	T	T	T	T	**S**
94	V	V	V	A	V	**T**	I	V	V	V	A	**W**
98	I	L	L	L	F	L	I	I	L	I	F	L
100	I	I	I	V	V	V	I	I	V	I	I	V
129	F	F	F	F	F	F	F	F	F	F	F	F
163	L	L	L	L	**F**	L	L	L	L	L	L	L
190	P	P	P	P	P	P	P	P	P	P	P	P
231	I	I	I	L	I	I	I	I	I	I	I	I
232	V	V	V	V	V	V	V	V	V	V	V	V
235	V	I	I	I	V	V	I	I	V	V	V	V
236	P	P	P	P	P	P	P	P	P	P	P	P
241	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
283	V	L	L	L	I	V	A	A	A	A	**K**	**K**
285	V	V	V	V	L	L	V	V	V	V	I	V
290	L	L	L	L	L	V	L	L	L	L	L	A

Table 3: *p*-nitrophenyl butyrate hydrolysis activity of secreted lipases using either the PrePro region of YlLip2 or the original PrePro region of each lipase.

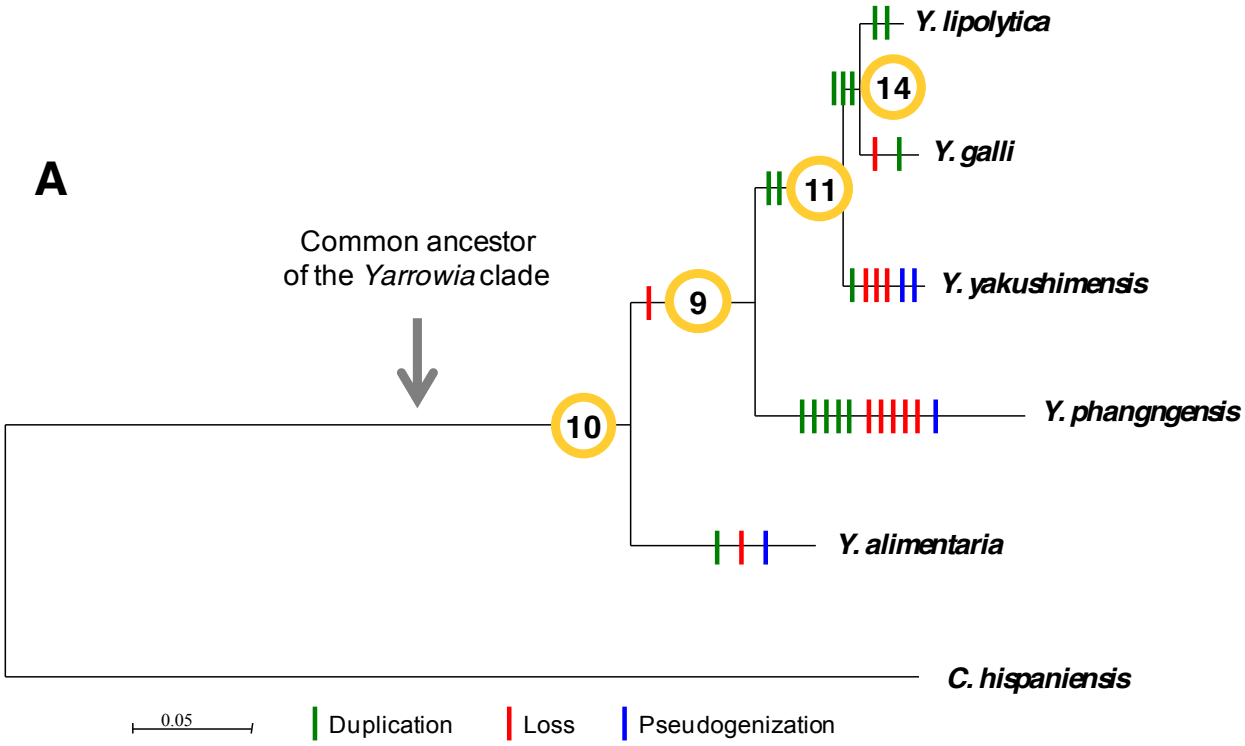
Enzyme	PrePro YlLip2*	PrePro YxLip2*
YlLip2	57.5	57.5
YpLip2a	19.4	23.7
YpLip2b	13.2	24.0
YpLip2c	23.3	25.7
YpLip2d	0	0
YpLip2e	0.2	0.9
YgLip2	16.8	12.9
YyLip2	4.3	5.8
YaLip2	58.7	50.6
YhLip2a	0.5	0.7
YhLip2b	0	0

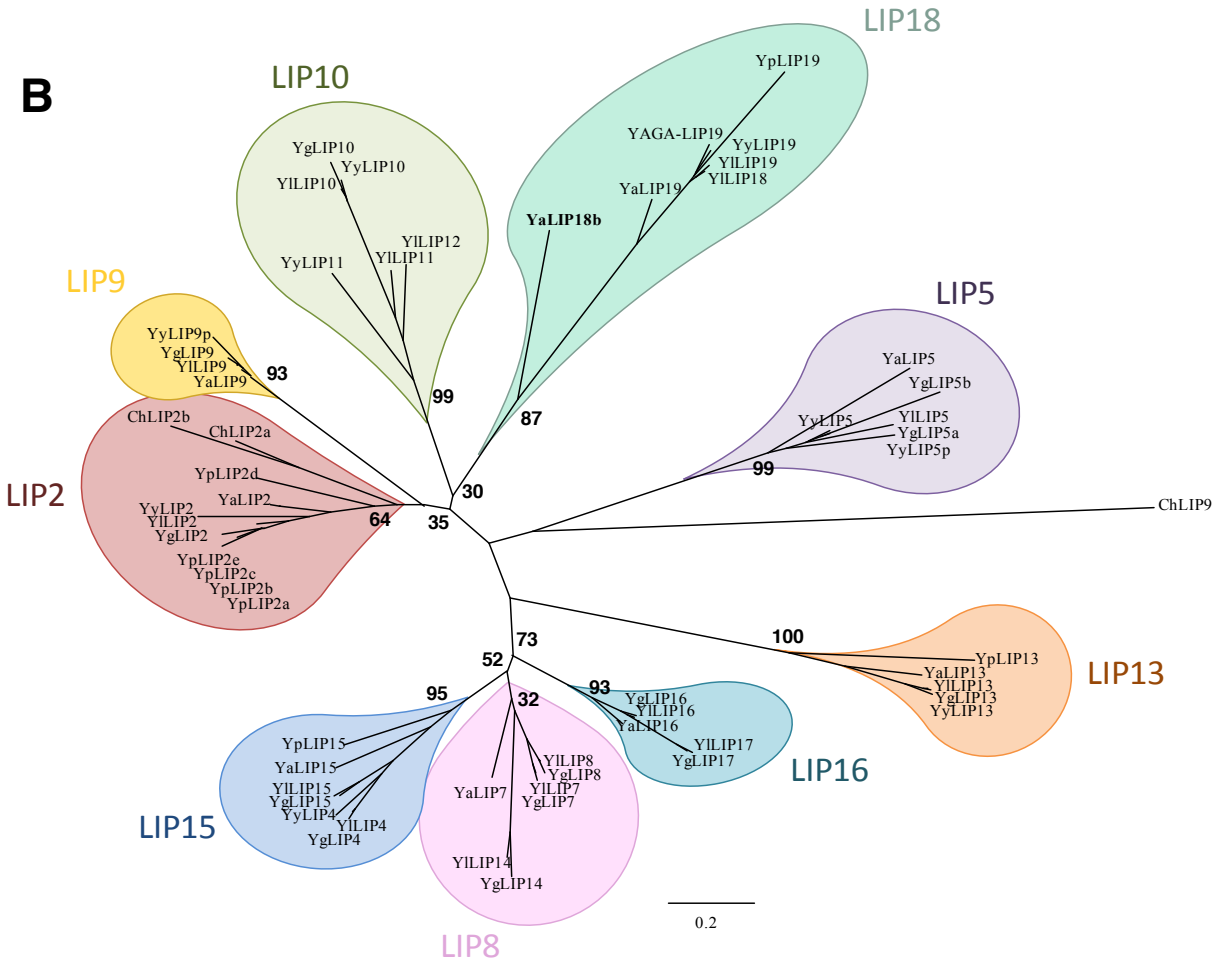
Experiments were performed in triplicate.
*µmol of *p*-nitrophenol liberated per minute.

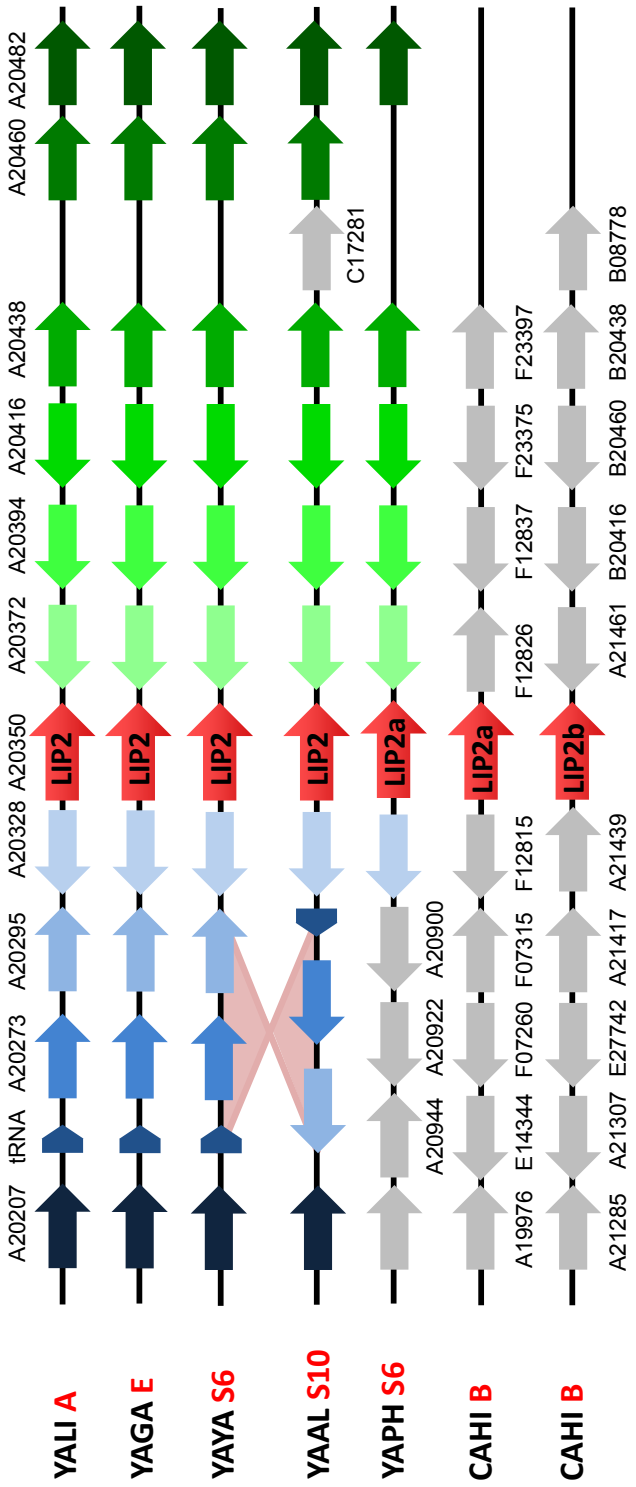
Table 4: 2-bromo-phenylacetic acid ethyl ester hydrolysis activity of wild-type lipases of the LIP2 family and of variants of YpLip2a

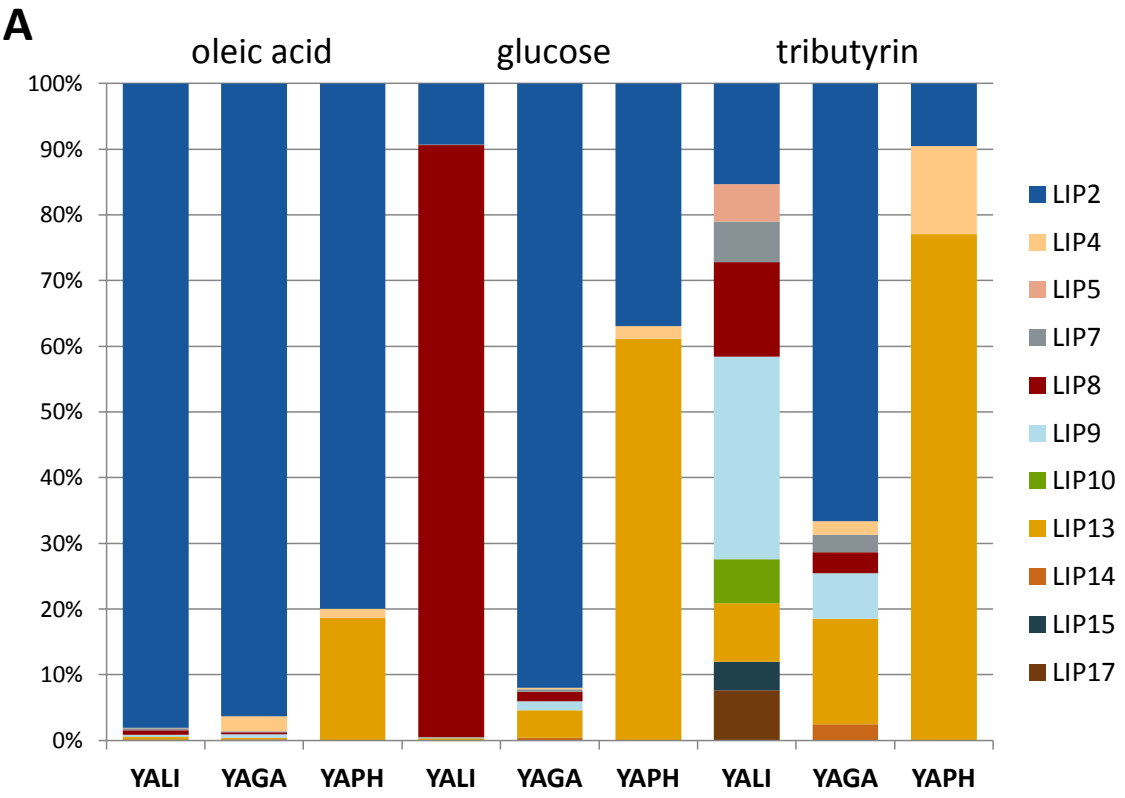
Enzyme	YlLip2	YpLip2a	YpLip2b	YpLip2c	YgLip2	YyLip 2	YaLip 2	YpLip2a V232S	YpLip2a D97A V232F
(S)-Initial rate[a]	2.4	4.6	1.02	1.29	2.41	3.2	2.16	11.6	0
(R)-Initial rate[a]	0.72	0.153	0.53	0.52	0.52	0.58	0.60	0	4.8
E-value [b]	3.3 (*S*)	30.1 (*S*)	1.9 (*S*)	2.5 (*S*)	4.6 (*S*)	5.5 (*S*)	3.6 (*S*)	> 200 (*S*)	> 200 (*R*)

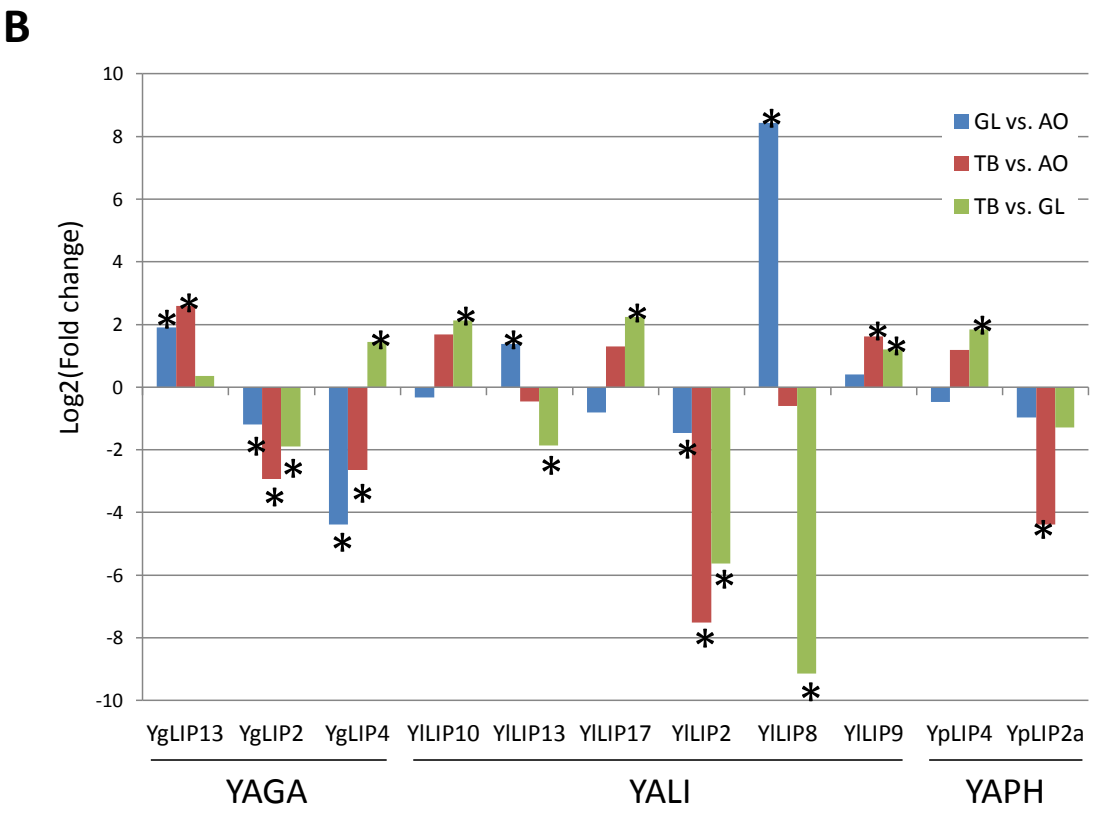
[a] µmol of 2-bromo-phenylacetic acid liberated per hour and mL of enzyme.
[b] E-value = viS/viR or viR/viS according to enantiomer preference; viR, viS: initial rates.

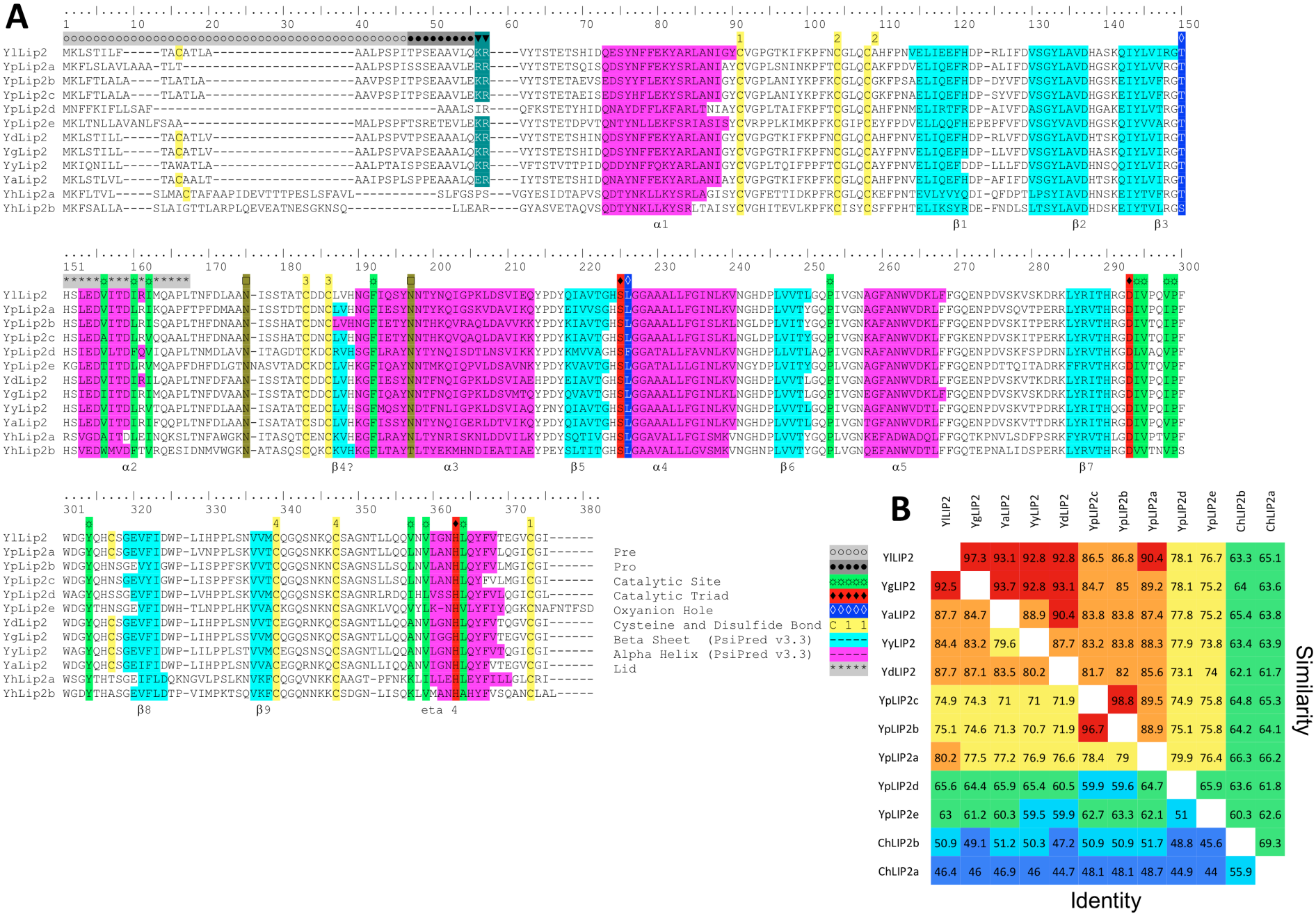

A
Common ancestor
of the Yarrowia clade
Y. lipolytica
14
Y. galli
11
Y. yakushimensis
9
Y. phangngensis
10
Y. alimentaria
C. hispaniensis
0.05
Duplication
Loss
Pseudogenization

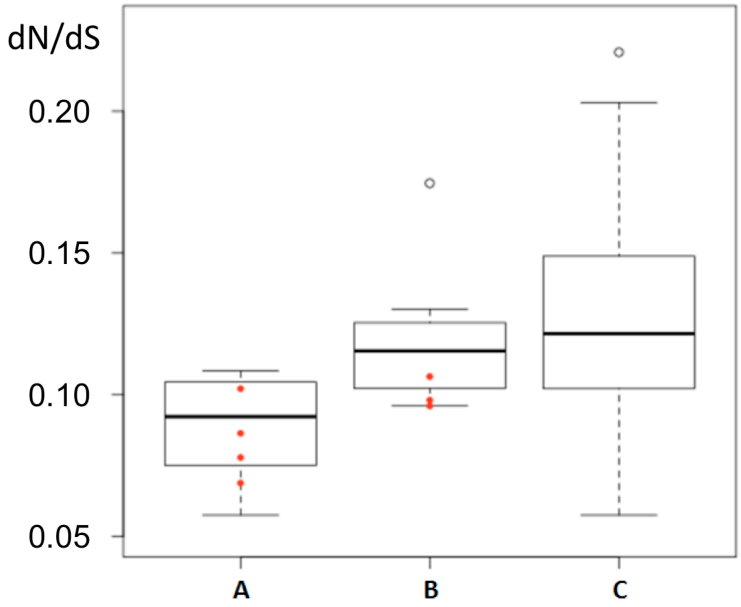

B
LIP18
LIP10
LIP9
LIP5
LIP2
LIP13
LIP16
LIP15
LIP8
YpLIP19
YAGA-LIP19
YyLIP19
YlLIP19
YlLIP18
YaLIP19
YaLIP18b
YgLIP10
YyLIP10
YlLIP10
YlLIP12
YlLIP11
YyLIP11
YyLIP9p
YgLIP9
YlLIP9
YaLIP9
YaLIP5
YgLIP5b
YyLIP5
YlLIP5
YgLIP5a
YyLIP5p
ChLIP2b
ChLIP2a
YpLIP2d
YaLIP2
YyLIP2
YlLIP2
YgLIP2
YpLIP2e
YpLIP2c
YpLIP2b
YpLIP2a
ChLIP9
YpLIP13
YaLIP13
YlLIP13
YgLIP13
YyLIP13
YgLIP16
YlLIP16
YaLIP16
YlLIP17
YgLIP17
YpLIP15
YaLIP15
YlLIP15
YgLIP15
YyLIP4
YlLIP4
YgLIP4
YaLIP7
YlLIP8
YgLIP8
YlLIP7
YgLIP7
YlLIP14
YgLIP14
93
99
87
30
35
64
99
100
73
52
95
32
93
0.2


YALI A
YAGA E
YAYA S6
YAAL S10
YAPH S6
CAHI B
CAHI B
A20207 tRNA A20273 A20295 A20328 A20350 A20372 A20394 A20416 A20438 A20460 A20482
A19976 E14344 F07260 F07315 F12815 F12826 F12837 F23375 F23397
A21285 A21307 E27742 A21417 A21439 A21461 B20416 B20460 B20438 B08778
LIP2
LIP2a
LIP2b
A20944 A20922 A20900
C17281

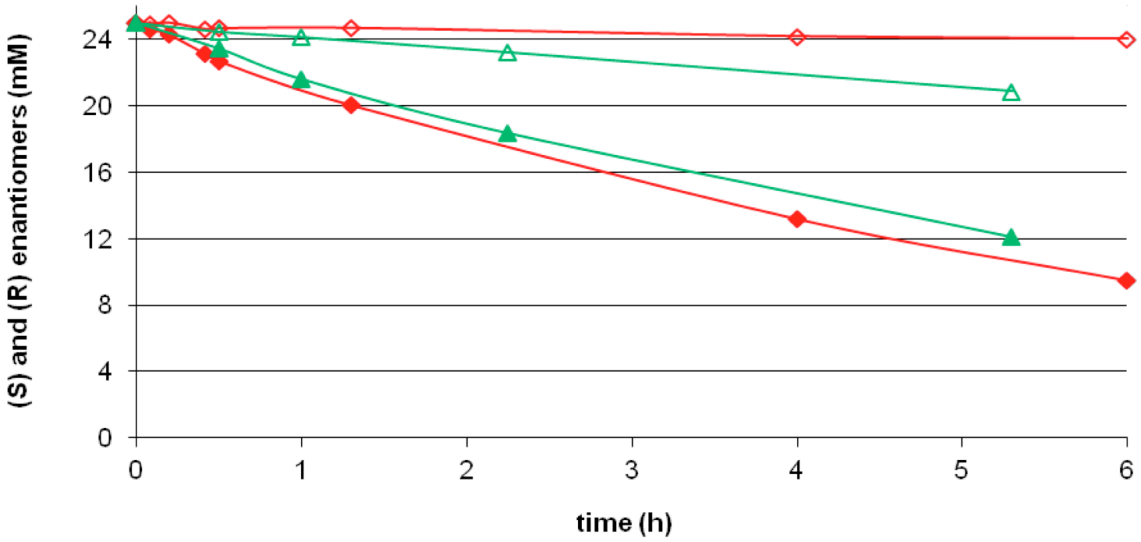

A
oleic acid
glucose
tributyrin
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
YALI
YAGA
YAPH
YALI
YAGA
YAPH
YALI
YAGA
YAPH
LIP2
LIP4
LIP5
LIP7
LIP8
LIP9
LIP10
LIP13
LIP14
LIP15
LIP17


B
Log2(Fold change)
10
8
6
4
2
0
-2
-4
-6
-8
-10
GL vs. AO
TB vs. AO
TB vs. GL
YgLIP13
YgLIP2
YgLIP4
YlLIP10
YlLIP13
YlLIP17
YlLIP2
YlLIP8
YlLIP9
YpLIP4
YpLIP2a
YAGA
YALI
YAPH

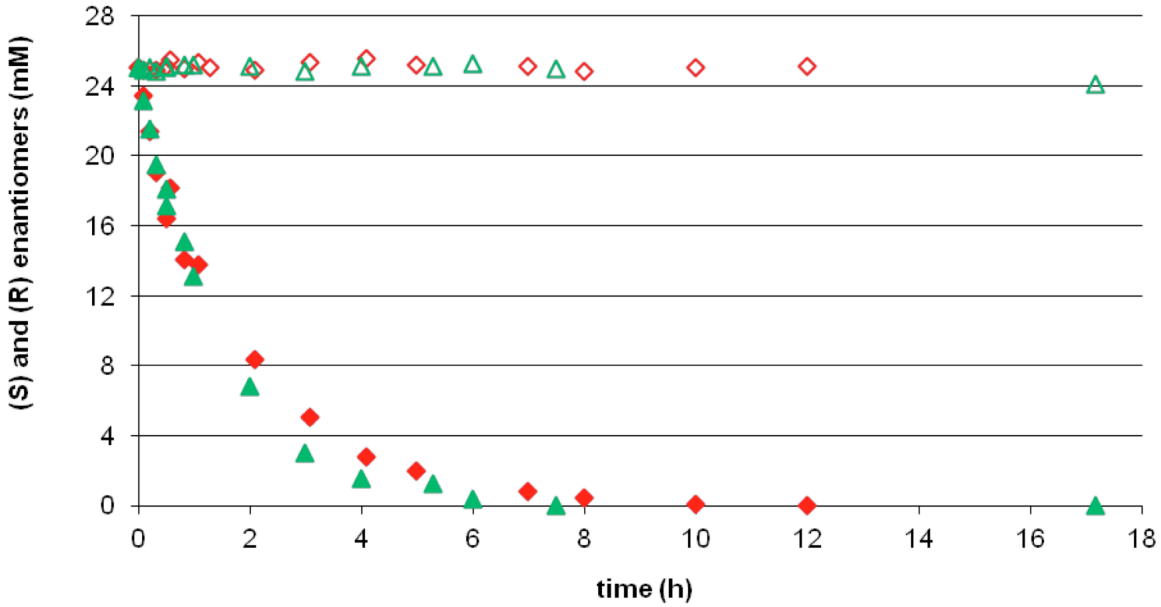

A
YlLip2
YpLip2a
YpLip2b
YpLip2c
YpLip2d
YpLip2e
YdLip2
YgLip2
YyLip2
YaLip2
YhLip2a
YhLip2b
α1
β1
β2
β3
α2
β4?
α3
β5
α4
β6
α5
β7
β8
β9
eta 4
Pre
Pro
Catalytic Site
Catalytic Triad
Oxyanion Hole
Cysteine and Disulfide Bond
Beta Sheet (PsiPred v3.3)
Alpha Helix (PsiPred v3.3)
Lid
B
YlLIP2
YgLIP2
YaLIP2
YyLIP2
YdLIP2
YpLIP2c
YpLIP2b
YpLIP2a
YpLIP2d
YpLIP2e
ChLIP2b
ChLIP2a
Identity
Similarity


dN/dS
0.20
0.15
0.10
0.05
A
B
C

A


(S) and (R) enantiomers (mM)
0
4
8
12
16
20
24
0
1
2
3
4
5
6
time (h)

B


(S) and (R) enantiomers (mM)
0
4
8
12
16
20
24
28
0
2
4
6
8
10
12
14
16
18
time (h)

Evolutionary genomics: an efficient tool to explore enzyme diversity and guide their engineering.

Applied Microbiology and Biotechnology

Muchalin Meunchan[1,2,3,4,#], Stéphanie Michely[5,6,#], Hugo Devillers[5,6], Jean-Marc Nicaud[5,6], Alain Marty[1,2,3,*], Cécile Neuvéglise[5,6,#]

Corresponding authors: Cécile Neuvéglise, INRA, ncecile@grignon.inra.fr
Alain Marty, INSA, marty@insa-toulouse.fr

[1] Université de Toulouse; INSA, UPS, INP; LISBP, 135 Avenue de Rangueil, F-31077 Toulouse, France
[2] INRA, UMR792 Ingénierie des Systèmes Biologiques et des Procédés, F-31400 Toulouse, France
[3] CNRS, UMR5504, F-31400 Toulouse, France
[4] Department of Biochemistry, Faculty of Science, Khon Kaen University, 123 Mittapap Road, Khon Kaen, 40002, Thailand
[5] INRA, UMR 1319 Micalis, F-78352 Jouy-en-Josas, France
[6] AgroParisTech, UMR Micalis, F-78352 Jouy-en-Josas, France

Supplementary data

Figure S1: Alignment of the PrePro regions of eight Lip2 lipases.

Figure S2: SDS-PAGE analysis of secreted LIP2 lipases.

Figure S3: Dotplot showing sequence similarity between YpLIP2b and pLIP2c from *Y. phangngensis,* within a 3-kb DNA region.

Table S1: Characteristics of the 61 lipase genes of the *Yarrowia* clade and *C. hispaniensis.*

Table S2: Expression data of the lipase genes in YALI, YAGA and YAPH in three media (oleic acid, glucose and tributyrin).

Table S3: Primers used for lipase cloning.

File S1: Amino acid sequences of the 59 full-length lipases of the *Yarrowia* clade and *C. hispaniensis*

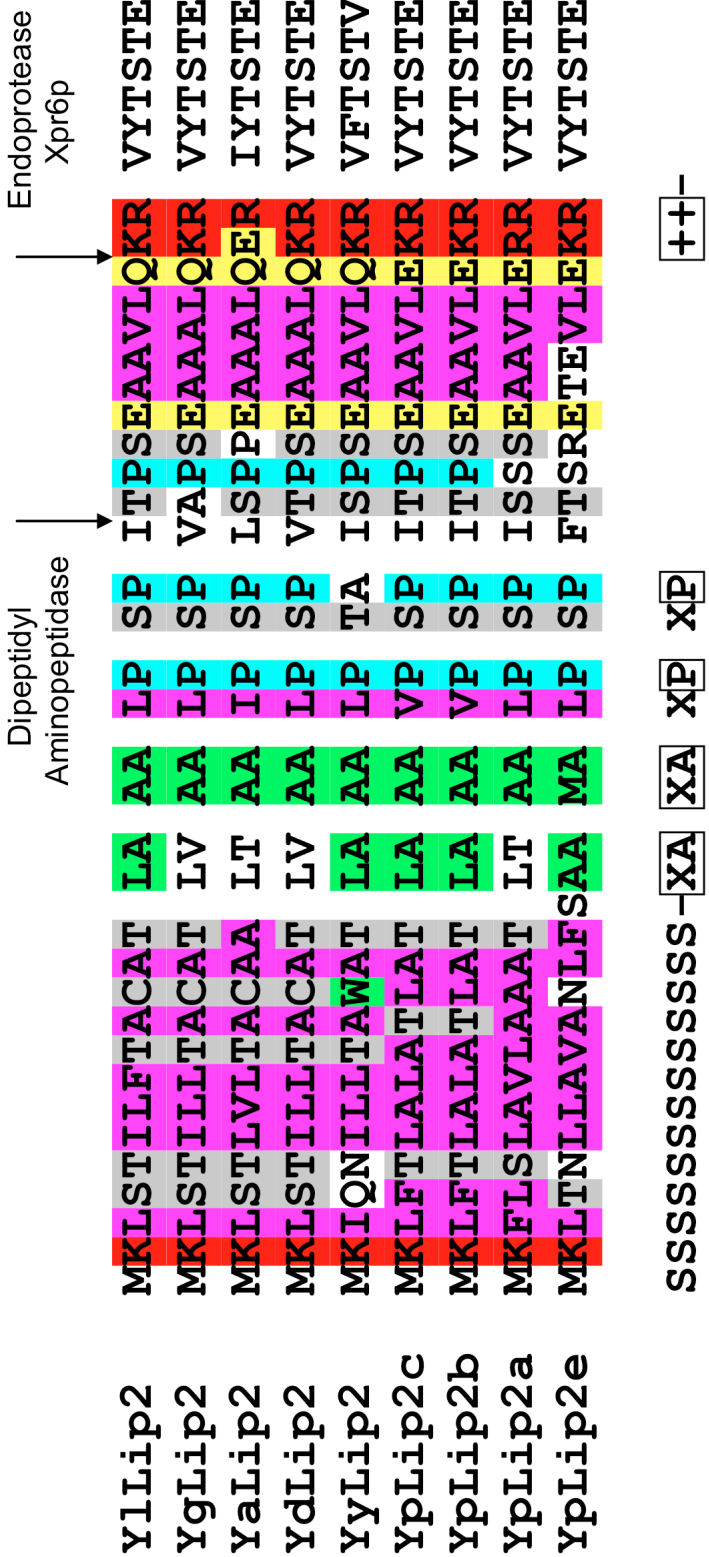


Fig. S1: Alignment of the PrePro regions of eight Lip2 lipases. Amino acids (aa) are colored according to their properties: positively charged aa in red, negatively charged aa in yellow, hydrophobic aa in pink, proline in cyan. Additionally, the X-Ala dipeptides are colored in green.

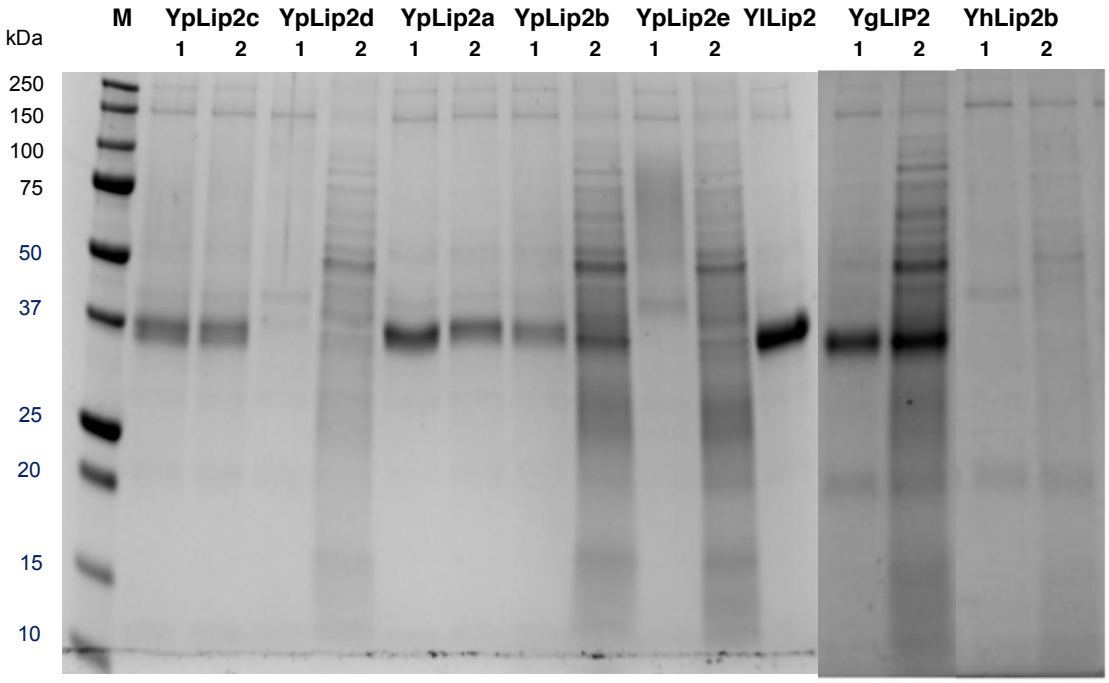


Fig. S2: SDS-PAGE analysis of secreted LIP2 lipases.
LIP2 genes were cloned in *Y. lipolytica* strain JMY1212 under the control of the constitutive TEF promoter from *Y. lipolytica*. Two strategies were adopted: either the entire genes were cloned with the presence of their own PrePro sequences (lines 1) or the part of the genes encoding the mature proteins were cloned behind the PrePro region of YlLip2 (lines 2).

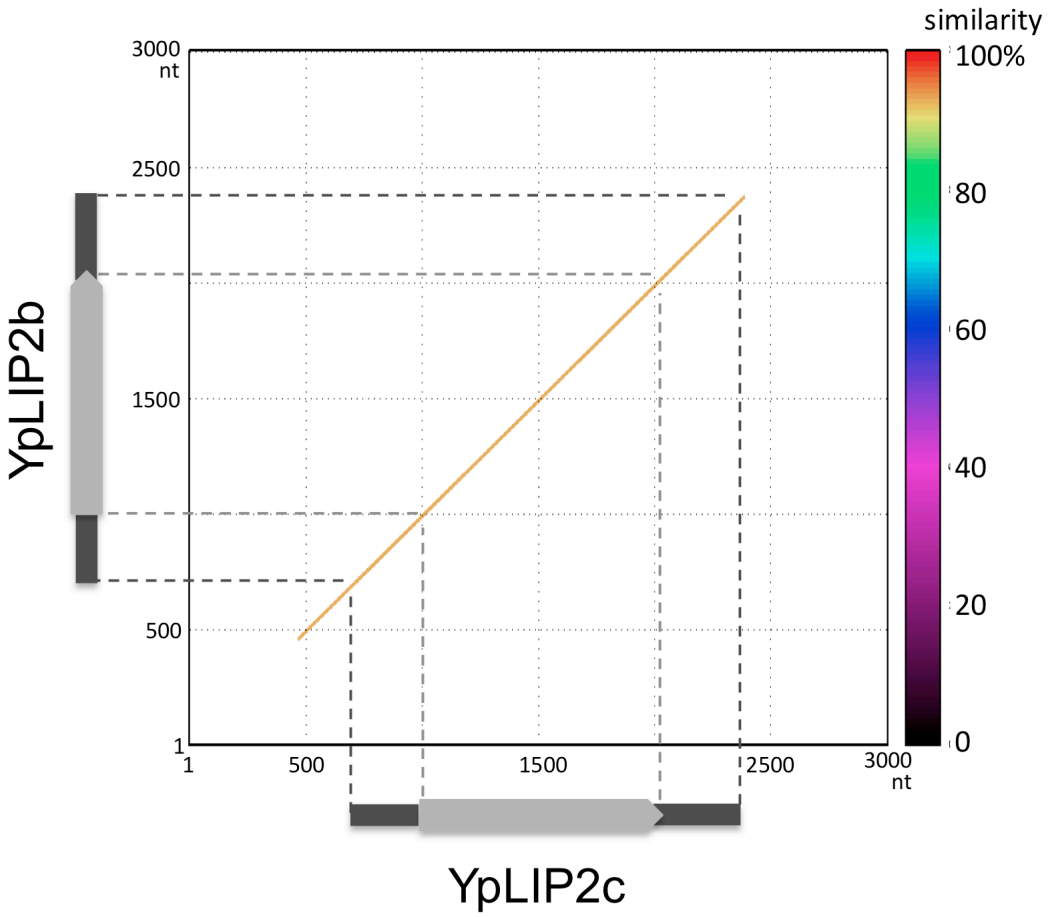


Fig. S3: Dotplot showing sequence similarity between YpLIP2b and YpLIP2c from *Y. phangngensis,* within a 3-kb DNA region (sizes are expressed in nucleotides, nt). The colour scale for similarity is given on the right. The CDS are depicted by grey arrows, the 5'- and 3'-UTR by dark grey rectangles. The coordinates of UTRs were deduced from the RNAseq reads obtained in the 6 samples (two replicates of 3 conditions: glucose, oleic acid and tributyrin). As shown by the orange line, the region of similarity exceeds the mRNA delimitation, suggesting that the duplication mechanism did probably not involve an RNA intermediate.

Table S1: Characteristics of the 61 lipase genes of the *Yarrowia* clade and *C. hispaniensis*

Species	Systematic Name	CDS length	Gene name	Remarks
Candida hispaniensis	CAHI0B09560g	343	LIP2a	
Candida hispaniensis	CAHI0C13190g	340	LIP2b	
Candida hispaniensis	CAHI0E00452g	299	LIP9	
Yarrowia alimentaria	YAAL0S10-01222g	334	LIP2	
Yarrowia alimentaria	YAAL0S02-06656g	204	LIP2p	pseudo: truncated copy
Yarrowia alimentaria	YAAL0S02-16094g	372	LIP5	
Yarrowia alimentaria	YAAL0S04-02696g	361	LIP7	
Yarrowia alimentaria	YAAL0S08-03510g	381	LIP9	
Yarrowia alimentaria	YAAL0S01-08218g	319	LIP13	
Yarrowia alimentaria	YAAL0S01-09208g	360	LIP15	
Yarrowia alimentaria	YAAL0S06-02102g	366	LIP16	
Yarrowia alimentaria	YAAL0S04-06656g	354	LIP18b	
Yarrowia alimentaria	YAAL0S01-04962g	390	LIP19	
Yarrowia galli	YAGA0E38754g	334	LIP2	
Yarrowia galli	YAGA0E25950g	406	LIP4	
Yarrowia galli	YAGA0E32462g	366	LIP5a	
Yarrowia galli	YAGA0F11166g	360	LIP5b	
Yarrowia galli	YAGA0D10484g	366	LIP7	
Yarrowia galli	YAGA0F22892g	370	LIP8	
Yarrowia galli	YAGA0E00738g	382	LIP9	
Yarrowia galli	YAGA0F31230g	338	LIP10	
Yarrowia galli	YAGA0E34992g	320	LIP13	
Yarrowia galli	YAGA0F20142g	352	LIP14	
Yarrowia galli	YAGA0E29558g	400	LIP15	
Yarrowia galli	YAGA0D12134g	364	LIP16	intron
Yarrowia galli	YAGA0F44122g	357	LIP17	intron
Yarrowia galli	YAGA0F28348g	379	LIP19	
Yarrowia lipolytica	YALI0A20350g	334	LIP2	
Yarrowia lipolytica	YALI0E08492g	406	LIP4	
Yarrowia lipolytica	YALI0E02640g	370	LIP5	
Yarrowia lipolytica	YALI0D19184g	366	LIP7	
Yarrowia lipolytica	YALI0B09361g	371	LIP8	
Yarrowia lipolytica	YALI0E34507g	412	LIP9	
Yarrowia lipolytica	YALI0F11429g	343	LIP10	
Yarrowia lipolytica	YALI0D09064g	429	LIP11	
Yarrowia lipolytica	YALI0D15906g	324	LIP12	
Yarrowia lipolytica	YALI0E00286g	317	LIP13	
Yarrowia lipolytica	YALI0B11858g	349	LIP14	
Yarrowia lipolytica	YALI0E11561g	406	LIP15	
Yarrowia lipolytica	YALI0D18480g	365	LIP16	intron
Yarrowia lipolytica	YALI0F32131g	363	LIP17	intron
Yarrowia lipolytica	YALI0B20350g	376	LIP18	
Yarrowia lipolytica	YALI0A10439g	377	LIP19	
Yarrowia phangngensis	YAPH0S6-00584g	334	LIP2a	
Yarrowia phangngensis	YAPH0S6-03488g	334	LIP2b	
Yarrowia phangngensis	YAPH0S1-03994g	334	LIP2c	
Yarrowia phangngensis	YAPH0S1-31032g	320	LIP2d	
Yarrowia phangngensis	YAPH0S5-13278g	342	LIP2e	
Yarrowia phangngensis	YAPH0S3-02454g	218	LIP2p	pseudo: truncated copy
Yarrowia phangngensis	YAPH0S2-22078g	304	LIP13	
Yarrowia phangngensis	YAPH0S3-09582g	357	LIP15	
Yarrowia phangngensis	YAPH0S5-11738g	368	LIP19	
Yarrowia yakushimensis	YAYA0S6-01420g	334	LIP2	
Yarrowia yakushimensis	YAYA0S2-08614g	400	LIP4	
Yarrowia yakushimensis	YAYA0S2-02982g	364	LIP5	
Yarrowia yakushimensis	YAYA0S4-04016g	378	LIP5p	pseudo: stop in frame
Yarrowia yakushimensis	YAYA0S2-34200g	383	LIP9p	pseudo: frameshift
Yarrowia yakushimensis	YAYA0S1-32616g	339	LIP10	
Yarrowia yakushimensis	YAYA0S1-12420g	325	LIP11	
Yarrowia yakushimensis	YAYA0S2-00584g	317	LIP13	
Yarrowia yakushimensis	YAYA0S7-05842g	374	LIP19	

Table S2: Expression data of the lipase genes in YALI, YAGA and YAPH in three media (oleic acid, glucose and tributyrin)

			Adjusted pValues			Read counts						log2(Fold change)		
Gene name	**lipase**	**CDS length**	**AO_GU**	**AO_TB**	**GU_TB**	**AO1**	**AO2**	**GU1**	**GU2**	**TB1**	**TB2**	**AO_GU**	**AO_TB**	**GU_TB**
YAGA0E34992g	LIP13	320	7.49E-05 *	4.85E-16 *	5.23E-01	12	17	239	98	132	109	1.906	2.594	0.349
YAGA0F20142g	LIP14	352	9.54E-01	4.55E-01	3.13E-01	5	19	26	12	12	28	-0.056	0.479	0.624
YAGA0E38754g	LIP2	334	8.06E-04 *	8.77E-25 *	4.96E-17 *	3183	8669	4438	3388	600	442	-1.197	-2.938	-1.896
YAGA0E25950g	LIP4	406	3.39E-39 *	1.96E-18 *	4.84E-03 *	140	210	10	7	25	14	-4.390	-2.654	1.440
YAGA0D10484g	LIP7	366	7.52E-01	1.16E-01	1.28E-01	8	9	22	18	30	15	0.248	0.856	0.794
YAGA0F22892g	LIP8	370	2.24E-01	5.48E-01	5.48E-01	17	26	86	51	26	30	0.657	0.343	-0.307
YAGA0E00738g	LIP9	382	9.75E-01	3.00E-01	1.37E-01	42	28	89	45	60	64	-0.029	0.563	0.672
YALI0F11429g	LIP10	343	5.95E-01	1.42E-03	3.39E-04 *	11	4	5	3	21	37	-0.330	1.677	2.123
YALI0E00286g	LIP13	317	9.38E-09 *	1.67E-01	4.46E-10	60	42	118	107	35	36	1.372	-0.462	-1.864
YALI0E11561g	LIP15	406	6.66E-01	2.54E-02	2.08E-02	9	6	7	2	13	31	-0.271	1.274	1.557
YALI0F32131g	LIP17	363	1.31E-01	2.49E-03	7.10E-05 *	15	11	4	5	26	44	-0.816	1.303	2.242
YALI0A20350g	LIP2	334	2.75E-06 *	3.32E-293 *	1.22E-47 *	14468	12569	2739	4245	67	62	-1.459	-7.509	-5.635
YALI0E02640g	LIP5	370	7.75E-01	2.99E-03	1.91E-02	7	7	9	5	17	36	0.179	1.598	1.374
YALI0D19184g	LIP7	366	3.75E-01	8.73E-02	5.82E-01	47	42	35	22	25	32	-0.324	-0.593	-0.246
YALI0B09361g	LIP8	371	0.00E+00 *	1.33E-02	5.75E-280 *	116	93	36218	39173	64	70	8.436	-0.597	-9.143
YALI0E34507g	LIP9	412	1.75E-01	1.21E-12 *	9.79E-07 *	63	40	65	48	139	181	0.406	1.620	1.198
YAPH0S2-22078g	LIP13	304	4.06E-01	3.49E-01	6.94E-01	98	180	165	114	92	99	0.387	0.257	-0.158
YAPH0S3-09582g	LIP4	357	5.36E-01	9.87E-03	6.13E-04 *	11	13	6	4	22	17	-0.467	1.192	1.837
YAPH0S6-00584g	LIP2a	334	9.36E-02	8.18E-42 *	6.40E-02	574	735	162	23	12	14	-0.965	-4.386	-1.294

The adjusted P-value cut-off for the differential expression was fixed to 1.e-3 in YALI and 5.e-3 in YAGA and YAPH.

* Differential expression is statistically significant

Table S3: Primers used for lipase cloning

Enzyme	PrePro region	Primer sequence
YlLip2a	PrePro YlLip2	GACGAGTGTACACCTCCACGGAGACTTCCCAGATCAG CCAAGGCATTTGCGGCATCTAACCATCCTAGGATGTAA
YlLip2a	PrePro YxLip2	CCGAAGGATCCCACAATGAAGTTTCTAAGTCTCGCTG ACCCTAGGTACCTTAGATGCCGCAAATGCCTT
YlLip2b	PrePro YlLip2	AACGAGTGTACACCTCTACTGAAACCGCTGAAGTCAGTG GTGCTCATGGGCATCTGTGGTATCTAACTTTCCTAGGATTGAT
YlLip2b	PrePro YxLip2	CCGAAGGATCCCACAATGAAGCTCTTCACACTCGCTCTTG ACCCTAGGTACCTTAGATACCACAGATGCCCATGAGCACAAAG
YlLip2c	PrePro YlLip2	AACGAGTGTACACCTCTACTGAAACCGCTGAAATCAGTG GTTCTCATGGGCATCTGTGGTATCTAACTTTCCTAGGATTGAT
YlLip2c	PrePro YxLip2	CCGAAGGATCCCACAATGAAGCTCTTCACGCTCGCTCTTG ACCCTAGGTACCTTAGATACCACAGATGCCCATGAGAACAAAG
YlLip2d	PrePro YlLip2	TCCGCCTGTACAAATCCACAGAAACCTATCACATTGATCAG GTTCTTCAGGGTATCTGCGGCTTGTAGTAGGCCTAGGAGATCT
YlLip2d	PrePro YxLip2	CCGAAGGATCCCACAATGAATTTCTTCAAGATCTTTCTACTTTCAGC ACCCTAGGTACCTTACAAGCCGCAGATACCC
YlLip2e	PrePro YlLip2	AGCGAGTGTACACATCCACAGAGACTGATCCTGTGACC GTAAGTGCAACGCTTTCAACACTTTTAGTGATTAGTTAACCTAGGATGAAT
YlLip2e	PrePro YxLip2	CCGAAGGATCCCACAATGAAATTGACTAACCTTCTTGCTG ACCCTAGGTACCTTAATCACTAAAAGTGTTGAAAGCGTTG
YgLip2	PrePro YlLip2	CCAGAAGCGAGTGTACACCTCTACCGAGACTTCC GATAAATACCCTAGGTTAGATACCACAGACACCCTC
YgLip2	PrePro YxLip2	CCGAAGGATCCCACAATGAAGCTTTCTACCATCCTTCTTACAGC ACCCTAGGTACCTTAGATACCACAGACACCCTCG
YyLip2	PrePro YlLip2	GAAGAGAGTGTACACTTCTACCGTGACCACTCCC GAAGAAGTGATACCTAGGTTAGATACCACAGATACC
YyLip2	PrePro YxLip2	CCGAAGGATCCCACAATGAAGATCCAAAACATCCTCCT ACCCTAGGTACCTTAGATACCACAGATACCCTGG
YaLip2	PrePro YlLip2	CAGGAGCGAATGTACACCTCTACCGAGACC GAAAGAGAAGCCTAGGGTAACTAAATACCACAAACACCTC
YaLip2	PrePro YxLip2	CCGAAGGATCCCACAATGAAGCTTTCCACCCTCGTCCTCAC ACCCTAGGTACCTTAAATACCACAAACACCTTCAGTGACAAAGTACTGGAG
YhLip2	PrePro YlLip2	AAGCTCGAGTGTACATTAGTGTAGAGACTGCTCAAG GGCACTATCCTAGGATGTCTTACAAAGCTAGGCAAT
YhLip2	PrePro YxLip2	ACCCTAGGTACCTTACAAAGCTAGGCAATTAGCCT CCGAAGGATCCCACAATGAAGTTCCTCACTGTTC
YlLip2d F163L		CACTCTCTGGGAGGAGCAACTGCCCTTTTG TCCTCCCAGAGAGTGCCCTGCAACAACCAT
YlLip2a D97A		ATCACTGCTCTTCGAATTAAGCAGGCTCCT TCGAAGAGCAGTGATGACATCTTCCAAGGA
YlLip2a V232F		GATATTTTCCCTCAGATTCCTTTCTGGGAT CTGAGGGAAAATATCTCCTCGGTGAGTAAC
YlLip2a V232S		GATATTTCCCCTCAGATTCCTTTCTGGGAT CTGAGGGGAAATATCTCCTCGGTGAGTAAC

Supplementary File S1: Amino acid sequences of the 59 full-length lipases of the Yarrowia clade and C. hispaniensis

>YAYA0S1-32616g LIP10
MIFQTLLVFSALLGLSIADPNAVLLGDAPPQRTYSDRVIQATQQTWDWMNYALHLSGLAS
CVGWQGIEYPFSCDSYCGDYPGMELIAEFGPKDIIDFSVSGFFSVDHTRKEIWHVFRGTV
SLTDAVSDVRLDTQPFDAWGPNHPNCSDCQAHAGFLKSYDLAYQQFEKNMTDAFQKYPDY
SLKVTGHSMGGAAAYVHGISMKTRGYDPYVVASGQPLVGNQALADYNDKLFFGDKPDFLR
QDEDRRYWRLTHKGDLVPQIPFWSPFHQPGGEIYIDYALADPPLDSLKVCDGQDNPNCNY
SSSMVYAALGGTLLWAHFQYFVIFTLCGVNYWSTHIHDG
>YAGA0F31230g LIP10
MKLPTLLLITTFLGLSTADPNAVLWGSAPAQKTYSDRVVQATQETWDWMNYALHLSGLAS
CVGFQGIQYPFTCDSYCADYPGMELIAEFGPKDIVDFSVSGFFAVDHHRKEIWHVFRGTV
SLTDGISDVRLDSLPFDAWGPNHPNCTDCKAHAGFLQSYKLAYAQFEKNMTEAFQKYPDY
SLKVTGHSMGGAASFIHGINMKTSGYDPYVVASGQPLVGNQALANYNDMLFFGDKPDFLR
QDKDRRYWRLTHKGDLVPQIPFWSPFHQPGGEIFIDYALADPPLDALKVCDGQDNPNCNY
SSNMINSAITGTVMWAHFQYFVVFTLCGVNFWSTHIHS
>YALI0F11429g LIP10
MKLSTLLLYTTLLLVGFSAADPNAVLLGSAPPQKKYSDRVVQATQATWDWMNYALHLSGL
ASCVGFQGIQTPFTCESFCADYPGMELIAEFGPRDVVDFSVSGFFAVDHSRKEIWHIFRG
TVSLTDGISDVRLDTLPFDAWGPNHANCSDCQAHAGFLQSYNLAYAQFEKNMTDTFKKYP
DYSLKVTGHSMGGAASFIHGINMKTRGYDPYVVASGQPLVGNQALADYNDRLFFGDKPDF
LRQDSGRRYWRLTHKGDLVPQIPFWSPFQQPGGEIYIDYVLSDPPLDSLKVCDGQDNPNC
NYSSNMVNSAITGTLLWAHFQYFVVFTLCGVNYWSTHIHG
>YAYA0S1-12420g LIP11
MKLLSVFSTLTSLLSLTTAQNAVLPADQATWDLLNRVMYVSGNAVCVGQQGIQYPFTCDA
FCSSFPDFELITSWDNLQPLEFDVSGFLAVDHKRKEFWHVFRGTNTLKEQITNLQVQQQP
MAKWYSPELAQCDNCTAHQGFQIVYHNAYELFGDKMKETWAKYPDYKNVVTGHSLGAAAA
YLHGLNLKTSGKDPLVITSGQPLVGNRALAEFNDRLFFGDKPDFTALGPSRQFYRLTHRE
DIIPRLPFWDPYYHSGGEIFIDYPWSQPPLEDLKVCDGAENPNCVFSTSILEDGTVGIVE
LAHLIYFINFSLQCAIPLSPAYTGN
>YALI0D09064g LIP11
MKASSMCSNQINQLAGFILDLHLSIFVLLMLQYSSETFLNRVVEGWSGAKGPENPTCKVI
KRSTRSRRNSPESWTVSFSFFFASSFCSFSCFAMLLSSLVLSALSLVSVTAQNVVQATQD
TWDLINYAEHLSSLAICGEPYGVIYKPFQCAGRCSDFPDMELITQFTPQDPLDFSVSGFL
AVDHKRKVFWHVFRGTATLNNGLTDLRIKRQPLTSWNTAKMDCPDCQVHVGFLQAYNLAY
SEAKGAMDDTFAKYPDYQVIVTGHSLGGAATFLHGINLKTSGYDPLVITSGQPLTGNKAL
ADYNDKLFFGDNPDFTHQGPDRRFYRVTHKEDIVPRIPFWTPYHQSGGEVYIDFPGINPP
VNTLKVCDGQQNPLCSFSTSLASTATQGIVEAAHLIYFTFFFLCSTLLYPPLNSDLPVGV
WGKPLNGTI
>YALI0D15906g LIP12
MRPSITIVGFLIASAVALGITQASKDTYNLLNYAENLNGVAVCVEQIGGIQKPFKCLGHC
DDFPDMELITTFHPKKLFDFSTTGFLAIDHKRKQFWHVFRGTASLTDGISNLRLERQPLV
FWDNPEFDCPGCEAHEGFLTAYNDAYDQIRDVLNQTLAQYPDYQIIVTGHSFGGASSFLH
GINLKSQGMDPLVITSGQPLTGNKALADFNDKLFFGDNPDFTYQGPDRRFYRVTHKDDLV
PRLPFWNPFHHSGGEVYIDYPLTNPPLRTLKICDGQQNPRCSFSTSLITAALLGTLQQAH
FMYFTFFSLVCGVNIAGHLGPPLQ
>YAYA0S2-00584g LIP13
MRLLTAIGFLLLLITQTQAISRATYELLQFHSQLSDIAYCVQTVRGPTQLRRPFKCGVQC
RKAQLEGVEVVHTFVHSPASPALTGYMAVDHTNQTKYVVFRGTNSLEDSMLDLSFGHEPT
ASPSFLKSTCPECKVQGAIMQAYDAFWAENSIQLSDFLYENFSQYSLSVTGHSLGGVAAA
LLATDLKLQGLDPLLINFGQPQYANLAYAQLVDSLFFPPADKEVIDPLYDSPQRRLYRVT
HWNDVFVTLPGQRGFFHSLGEVYISYPTVNPPRRTIRYCEGPESEYCHSGDYNPFERANF
LKNHLAYFGWVGYCPYL
>YAPH0S2-22078g LIP13
MILASTTIFLWIFQLILAQTVSQQTYDLLKFHARLSDIAYCVQPVNSPTQLRQPFECGVK
CTNIEGVELFHSFVHKPGNTGFTGYFAFDHVNKTKYIVFRGTNSFEDTLVDFSMSHELPC

ETPFFLENLCPECKVQSSLLEVYGLFLEQHVTPLWDFVYTFYPDYSISVTGHSLGGVAAA
FLATHLRVQGLEPVLVTFGQPRYGNLAHAEFIEKIFLTSQRRLYRVTHWNDVFVSLPSSE
QFSHFSGEIYISHPWVDPPLETVRYCEGRESETCHAGDYNPVERVTFLKNHLAYFIWVGY
CPYR
>YAAL0S01-08218g LIP13
MWITVLPILFFLPALVLAISAPTYDLLAWHSQLSDVAYCVQTVRGPTQLRTPFKCGKQCR
EPQMQDMQLIHAFVHSPARPALTGYLAVDHASHTKYVVFRGTNSIEDTVLDLSFGQQPSV
SPGFLQGACPECQVQGGILAAYETLWEENSGPLSDFLHSQYPNYTLSVTGHSLGGVAAAL
LATHAKLEGLDPTLVTFGQPRYANLAYAQLVDSLFFPLNGSDGGSNLGPLDESPRRRMFR
VTHWNDVFVSQPELGGFHHSLGEVYISYPWVSPPRRTVRYCDGAQSEYCHRGDYNPLERA
NFLKNHLAYFGWIGYCPYL
>YAGA0E34992g LIP13
MLLRILPTTLVLLFFLVTQTHAISEATYELLAFHSQLSDIAYCVQTVRSPTQLRHPFKCG
VQCRKTQFEGLEVIHTFVHKPASPALTGYMAVDHTNQTKYVVFRGTNSLEDSMLDLSFGY
EPTASPSFLKDACPECKVQGAILEAYEAFWAENCDQLSEFLYEDYPHYSLSLTGHSLGGV
AAALLATDIKVQGLDPLLINFGQPQYANLAYAQLVDSLFFPQTDKEVIDPLYDSPQRRLY
RVTHWNDLFVSLPGQRGFFHSLGEVYISYPWVKPPRRTIRYCEGPQSEYCHGGDYNPLER
ANFLKNHLAYFGWIGYCPYL
>YALI0E00286g LIP13
MRPFTTLVLLLLLITQTCAISEATYELLQFHSQLSDIAYCVQTMRGPTQLRHPFKCGVQC
RKTQFEGLEVVHTFVHKPARPALTGYMAVDHTNQTKYVVFRGTNSLEDSMLDLSFGYEPT
ASPSFLNDTCPECKVQRAILEAYDAFWAENCDQLSDFLYEDYPHYSLSLTGHSLGGVAAA
LLATDIKLQGLDPLLINFGQPQYANLAYAQLVDSLFFPQTDKEVIDPLYDSPQRRLYRVT
HWNDLFVSLPGQRGFFHSLGEVYISYPWVKPPRRTIRYCEGPQSEYCHGGDYNPLERANF
LKNHLAYFGWIGYCPYL
>YAGA0F20142g LIP14
MVLASSTVFAEWFFRVLLGTVAPSPQTATAPVTQEFYDTALTYSHLSNIAYCINAPFESL
KTDFSCGVACSHFPNMELVEAFGGEFFETSITGYLAIDHVKKEKYVVYRGTFDIGDVYTD
IQLAQSPYLVTASMFDSPDNLCEGCTIHDGFNKAYKETMVNIGDKLEQHLTNNSDYKLVV
AGHSLGAAIAVLSATSIKAKGFDPYLFTYGQPRIGNANFAQFVSKMWFGDGDGLSMGPDR
RLFRISHWNDLFVGFPAFKDYVHSVGEIYIDYYTVNPPLNTLYSCAGPESMSCYRKDFNA
IERANILKNHLAYIDWLSLCTLNVGRRELQERGRRFEGKFLYGGIANGTIIF
>YALI0B11858g LIP14
MVLSTVIGEWFSRVLFGTVAPSPLTATAPISQDFYDTALTFSHLSNVAYCINTPLESLKS
DFSCGVACSHFPNMELVEVFGGEFFETSITGFLSIDHVKKEKYVVYRGTYDIGDVYTDIQ
LSQSPFLVTPSALGSTANLCEGCTIHDGWNKAYNETMGIIGDKLADHVNSNPDYRLVVTG
HSLGAAIAVLSATSLKVNGQDPYLYTYGQPRIGNANFANFVSKQWFGEGDGLSMDSDRRY
FRLTHWNDLFVGFPAFKDYVHSVGEIYIDYFTVQPPLNKVFSCAGPESMSCYRKDFNALA
RLDIVKNHLAYFDWISLCTLNIGRRDLERGRKFEGTWLYGGLANGSTIF
>YAPH0S3-09582g LIP15
MVSWYTQFKAAILSLLGMTASTPSTTIASISQGMYNEISYFSRMVNTAYCANAPITPLRT
DFSCGDSCQYFANLKLDSIFGGNFYSTSATGLLAHDHKRKEKYVLFRGTFSIPDAYTDIQ
FQKSPWLAKLPNGVVPKIQSAGGQPLTCEGCAVHDGFAKAFNETLKNSGHQFDKFLANHT
DYKLYVVGHSLGGAMAQMFAVRLKLMGYDPTLITYGQPRVGNKEYAEFVSRLFFNDESGL
LVDENRRLYRVTHWNDIVVGLPNFADYTHSVGEVFIASEDLNPPVESVTLCEGAENEACH
RGHFSLWERVKILHNHLAYINYIGYCALNIGRRSILNMPNYRGKNTYAHKSETTKDE
>YAAL0S01-09208g LIP15
MQFTFGQVMALLYSTLFGRPPASKETTRVYVDQSMYDFTAKFSRLCNVAYCVDAPITPLR
TDFSCGESCRYFPNMTLDAVFGGDFYSTSITGYTGTDHHLKEKYIVFRGTFSIPDIVTDI
QFQRSPWLVQLPPIVTTKYNDFKPAARANCTDCQIHDGFAKAFNETLQNAGPQINDFLSN
NTDYKLYVAGHSLGAAQAQLFATSFKLQGYDPILVSYGQPRIGNKEFADFVSELFFDGDD
GLSMNDTRRMYRFTHWNDIFVGLPDWGNYTHSVGEIYIDQRDVYPPLNAVTACAGPENPD
CHRGTFNLWEQINLFQNHLAYIYYIGYCALNIGKRDVFNMPKYTGNYTYGHRSEYQTNLP
>YAGA0E29558g LIP15
MNLTLGQVVAYLYASLFGEPPATLSKTKVQASQDLYNFTAKFSRLANIAYCVNAPVTPLR
TDFTCGESCRYFPNMTLDSVFGGDFYSTSITGYIAYNHAEKEKYVVIRGTFSIPDAVTDI
QFQNAPWLVQLPEHLIPTQTDMQQKWAVRPDLEVENKGLDSLKERTPLTEDPRLIPIKTK

ECKGCMIHDGFAKSFNETMANAGPQFEKFLTNHTDYKMFITGHSLGAAQALLFGTHFKLL
GHDPTMINFGQPRVGNSEFANFINQLWFNDTGLIVDDKRRNYRITHWNDIVVGVPDWMNY
THSIGEVYIDYEDVNPPLNKVAVCEGGENEACHRGTFNLWSRINLLQNHLAYIYYIGYCA
LNIGRRDVMAMPKYHGNFSYRYGSDPGYSGNGRTELPTTN
>YALI0E11561g LIP15
MTFTIGIMNLTLGQVVAYLYASLFGEPPATLSKTRVQASQELYNFTAKFSRIANIAYCVN
APVTPLRTDFTCGESCRYFPNMTLDSVFGGDFYSTSITGYIAYDHLNKEKYVVIRGTFSI
PDAVTDIQFQQSPWLVELPKHLIPTQDDMKQRWAVRPDLEVENKGLDNLKERALVVEDPR
LIPIKTSECKGCMIHDGFAKAFNETMVNAAPQFEKFLTNHTDYKMYVTGHSLGAAQALLF
ATHFKLLGHDPTMINFGQPRVGNSEFANYINQLWFNDTGLEVNDKRRNYRLTHWNDIFVG
LPDWSNYTHAIGEVYIDQESVYPALDKVSVCEGGENEACHRGTFNLWSRIDLLQNHLAYI
YYIGYCALNIGRRNVLAMPKYHGNNSYKYATEPGYTGPGRTLVPTN
>YAAL0S06-02102g LIP16
MKLIAFILVLAGMLSAVHAQTKAITQDTYDLILKYGQLCNVAYCVRAPGPFGLQDNFTCG
KACSHFPDTELVYKFGGNFFSTSITGFLAVNHEKKEKYIVFRGTFSIADAITDIQFLQKP
YLADLPPLNTTNINSTHPLARPECPGCEVFDGFQKAYRETMVNMGDNLINHINSNPDYKL
IVTGHSLGAATGLLMAINLKNLGLDPVVVLYGQPRVGNKAFADYANSLFFDKGSNGVDIT
NSTRLYRVTHWNDVVVGLPFWSGYTHTIGEVYISYPNVASPIRYVHSCAGASNRACHSGN
FNLFARFNILKNHCSYLSWIFYCAINVGKRDLIHDPPRVSSGMKHWSEGQFSDQSEQQIY
DAVFPN
>YAGA0D12134g LIP16 [intron]
MLSLITLFLFIAQAYAQIANITQPTYDFILKYGQLSNVAYCVRALGPFDLTDQFTCGKSC
AHFPNMTLDYQFGGSFFSTSITGFLAHDHTKKEKYIVFRGTFSLADAYTDALFFQEPYMA
DLPPLNTTNINSTSTEARADCPGCEIHDGFQVAYRETMENMQDHLVDFLKNNTDYKLIVT
GHSLGAATSLLMAINLKNLGFDPMVITFGQPRVGNKAFADYANSLFFKEGDNGMDINPER
RLYRVTHWNDIVVGVPFWSGYTHTLGEVYISYPDVGSPIEYVNACAGPDNSQCHYGTFDL
LARVNILKNHCSYLNWIFYCAFNVDKRDMMIDPPRIDKRVQHWSGKFADVEVGERMMYEA
SYPM
>YALI0D18480g LIP16 [intron]
MLSLIAIFLLVTTALAQTANITQSTYDFVLKYGRLSNVAYCVKAPGPYELETDFTCGRSC
GHFPNVTLEHQFGGDFFSTSITGFLAHDHTKKEKYIVFRGTFSLADAITDALFLQEPYLA
DLPPLNTTNINSTSNSARVDCPDCEIHDGFQKAYRETMVNMQGHLVAFLRNNTDYKLIVT
GHSLGAATALLMGINLKNLGFDPMVITFGQPRVGNKAFADYADSLFFKQGDNGLNINPER
RLYRVTHWNDIVVGVPFWSGYTHTLGEVYISYPDGVNAPIEYVNACAGPDNDQCHYGSFD
LLARVNILKNHCAYLNWIFYCAFNVDKRQMMIDPPRIHKRVEHWSGKFADVEFSERMIYE
ATYPM
>YAGA0F44122g LIP17 [intron]
MLSFITIFLLATQALAQTAAISQSTYDFVLKYGWLSNVAYCVRAPGPFALQDNFTCGKSC
AHFPNVTLDYQFGGNFFSTSVTGFLAHDHTKKEKYIVFRGTFSLADAITDIQTVQQPYMS
TPPPLNTTNINSTDPSANIDCPGCEVHDGFQKAYRETMTNVQDHLVEFLGKNKDYKLIVT
GHSLGAVTALFMGINLKNLGYDPTMINYGQPRLGNKAFADYVDALFFKKGDDGLAITPER
RMYRVTHWNDFFVGWPAGYTHTIGEVYISDPTGINAPLQDVYACAGPENDKCHHGSFNIL
ERLNILKNHCAYLNWIFYCAINVDKRQMMIDPPRWTAGNSDGVELTERMMYEATYPM
>YALI0F32131g LIP17 [intron]
MLSFIALFLLVAQALAQTAPITQETYDFVLKYGWLSNVAYCVRAPGPFALQSDFTCGNSC
AHFPDVTLDYQFGGNFFSTSVTGFLAHDHTKKEKYIVFRGTFSIADAITDIQTIQQPYMT
SIPPLNTTDINSTNPSASINCPGCQVHDGFQKAYRETMVNVQDRLVDFLGNNTDYKLIVT
GHSLGAVTALFMGINLKNLGYDPTLINYGQPRLGNKAFADYVDALFFKQGDDGLTINPER
RMYRVTHWNDFFVGWPAGYSHTLGEVYISDPTGINAPIEDVYSCAGPENNQCHHGSFNLL
ERLNILKNHCGYLNWIFYCAINVDKREMMIDPPRVGKRVEHWSGKFSDVESTEGLMYEAI
YPM
>YALI0B20350g LIP18
MIVLLLLLSVALSSPIMLDRFGDSVNRMIRGNSEAYESGETRGVSQEFEQRLVRYMWFNN
VAPCVPKKLQHPFKCIARGCKELGKHTELVDIFTHSDNLFDRTISGFVALDHKHKEIVLA
LRGTQDAHDWVTDLHLRLVGLHPEHLGVSNFNCRNCQVDLGFLKGYLHSFHVVDSIVQRL
TEKYPDYQLVITGHSLGGTAATLFGLNYRLNGYSPLVFSAGAPALGNKQFANFADRVFWG
SQNPNTLKVKERDIKFCRMTHLGDFVPRFPFWNGYQQMSGEVFINDVRGIYPPRETLQRC

NGQQNRQCSFGDQYRKLEMNFKPHSAYLVPGSECSLSGRRELTFGQVHAANDTNKDSDTN
IEPDIEPILIPVVQID
>YAYA0S7-05842g LIP19
MILLFLFLCFALSSPIGMDRFSDGVNRMIRGNAGAFESSETLGVSREFEQRLIRYMWFNN
VAPCVPKKLQHPFKCIARGCKELGRQTELVDIFTHSRNLFDRTISGFVALDHKHKEILLV
LRGTQDINDWVTDLHLRLVDLQPEHLGLTNLNCRNCQVDQGFLKGYLHSFHAVDSIVRRL
IEKYPKYQLVITGHSLGGTAATLFGLHYQLHGSSPQVFSAGAPVLGNKQFANFADRVFWG
SQNPNTLIVKENDIRFCRITHLGDFVPRFPFWKGYQQMSGEIFINDVRGIEPPRDALKRC
NGQQNDKCSSGDQYRHLEMNFKPHSAYLVPGSECSFSGKGELTLGQIQAAKNTYQDTDTD
TDIEAVLIPVVQLD
>YAPH0S5-11738g LIP19
MLIIFLLTLIASASPFIRLALGEKLNQLLVGSHEAFESEETQLISPEFQECLIRYMWFNN
VAPCVPKQLEHPFVCAARGCKMLGKYTELVDIFTHSESRMDMLTDRTITGFVALDHLHKE
IVLVLRGTQDVNDWLTNLQVSLVGLNPADLGVKSLNCPNCRVGLGVLSGYISSQRVADPI
IRRLKEKFPQYQLVVTGHSLGGTAATLFGLNYQLNGLSPKVFSVGAPALGNKEFTNFVDE
VFWGSEHPDTLSVPEKNINFTRVTHLGDFIPRFPFWKGYQQMSGELFIHDVKGIDPPLDK
LRRCNGQQNRKCSFGDKFRQLTLNIRSHSAYLVPRNRCSGERRKIMGDIGAYGNDTVLIP
VVDLDSLE
>YAAL0S01-04962g LIP19
MILFFFLLICVGLSSPLLNRVGTGVNRMIGGSDEAYQSHETVAVSSGLEKRMTRYMWYNT
VAPCVPKRLQHPFKCEAQGCTKVGTQTELVDVFTHSGDVLDVFTDRTISGFLALDHKNKE
ILLVLRGTQDANDWVTDLRLRLVPLEASHLDAPSIGCHGCQVNVGFLEAYQHTWRVVDST
IRDLKKRYPNYQLVLTGHSLGGTAAILFGLNYKLNGLHPTVFTAGAPAIGNKQFANFADQ
VFWGSSNPDTLSVPENRLCFLRLTHRGDLIPRFPFWGGYQQMSGEIFINDIRGIYPPLSS
LRRCNGQQNNRCSFGDQNRQLEMNFRPHSAYLVPGSECSFSNREGHIGRGGAGDMVNAMD
TANTTEDTTDTTEMDPGLLAVEIPVVAGPV
>YAGA0F28348g LIP19
MIVLFLFLCVALSSPIVLDRFGDGVNRMIRGNSEAFESGETRGVSRYFEQRLIRYMWFNN
VAPCVPKKLEHPFKCIARGCKELGKHTELVDIFTHSDTLFDRTISGFVALDHKHKELVLV
LRGTQDANDWVTDLHLRLVELRPEHLGVSSFSCRNCQVDLGFLKGYLHSFHAVDSMVRRL
TELYPKYRLVITGHSLGGTAATLFGLNYQLHGFSPLVFTAGAPALGNRHFSNFADRVFWG
SQNPNTLKVKEGDIKFCRMTHLGDFVPRFPFWSGYQQMSGEVFINDVRGIDPPRESLQRC
NGQQNKKCSFGDQYRQLEMNFKPHSAYLVPGSECSLSGRRELTVGQAESEVQAAAATNNT
DTYADTDFEPVLISVLQID
>YALI0A10439g LIP19
MIILLLFLSVALSSPIMLDRFGDSMNSMIRGNNEAYESGETCGVSQEFEQRLVRYMWFNN
VAPCVPEKLQHPFKCIARGCKELGKHTELVDIFTHSDNLFDRTISGFVALDHKHKEIILA
LRGTQDVNDWVTDLHLRLVELHPEHLGVSNFNCRNCQIDLGFLKGYLHSFPAVDSIVQRL
TEKYPNYQLVITGHSLGGTAATLFGLNYRLNGYSPLVFSTGAPALGNKQFANFADRVFWG
SQNPNTLKVKERDIKFCRMTHLGDFVPRFPFWNGYQQMSGEVFINDVRGIYPPRETLQRC
NGQQNRQCSFGDQYRKLEMNFKPHSAYLVPGSKCSLSGGRELTFGEVHSAANDTNKDSDT
DIEPDIEPILIPVVQID
>OLHI0B09560g LIP2a
MKFLTVLSLMACTAFAAPIDEVTTTPESLSFAVLSLFGSPSVGYESIDTAPVSQDTYNKL
LKYSRLAGISYCVGFETTIDKPFRCGLQCKEFPNTEVLYVYQDIQFDPTLPSYIAVDHNS
KEIYTVFRGTRSVGDAITDLEINQKSLTNFAWGKNITASQTCENCKVHEGFLRAYNLTYN
RISKNLDDVILKYPDYSQTIVGHSLGGAVALLFGISMKVNGHDPLVVTYGQPLVGNKEFA
DWADQLFFGQTKPNVLSDFPSRKFYRVTHLGDIVPTVPFWSGYTHTSGEIFLDQKNGVLP
SLKNVKFCQGQVNKKCAAGTPFNKKLILLEHLEYFILLGLCRI
>OLHI0C13190g LIP2b
MKFSALLASLAIGTTLARPLQEVEATNESGKNSQLLEARGYASVETAQVSQDTYNKLLKY
SRLTAISYCVGHITEVLKPFKCISYCSFFPHTELIKSYRDEFNDLSLTSYLAVDHDSKEI
YTVLRGSHSVEDWMVDFTVRQESIDNMVWGKNATASQSCQKCKVHKGFLTAYTLTYEKMH
NDIEATIAEYPEYSLTITGHSLGAAVALLLGVSMKVNGHNPLVVTYGQPLVGNQEFANWV
DTLFFGQTEPNALIDSPERKLYRVTHRGDVVTNVPSWGDYTHASGEVFLDTPVIMPKTSQ
VKFCQGQNNKKCSDGNLISQKLVMANHAHYFVSQANCLAL
>YAYA0S6-01420g LIP2

MKIQNILLTAWATLAAALPTAISPSEAAVLQKRVFTSTVTTPIDQDDYNFFQKYARLANI
GYCVGPLTQIFPPFTCGLQCAYFPNVELIQEFDDPLLLFDVSGYLAVDHNSQQIYLVIRG
THSLEDVITDLRVTQAPLTNFDLAANISATATCEDCLVHSGFMQSYNDTFNLIGPKLDSV
IAQYPNYQIAVTGHSLGGAAALLFGINLKVNGHDPLVVTLGQPIVGNAGFANWVDTLFFG
QENPDVSKVTPDRKLYRITHQGDIVPQIPFWAGYQHCSGEVFIDWPLILPPLSTVVMCEG
QSNSQCSAGNTLLQQANVLGNHLQYFVTQGICGI
>YAPH0S1-03994g LIP2c
MKLFTLALATLATLAAAVPSPITPSEAAVLEKRVYTSTETAEISEDSYHFLEKYSRLANI
GYCVGPLANIYKPFKCGLQCGKFPNAELIQEFHDPSYVFDVSGYLAVDHGSKEIYLVFRG
THSLEDAITDLRVQQAALTHFDNAANISSHATCDNCLVHNGFIETYNNTHKQVQAQLDAV
IKKYPDYKIAVTGHSLGGAAALLFGISLKLNGLDPLVITYGQPIVGNKAFANWVDKLFFG
KENPDVSKITPERKLYRVTHRGDIVTQIPFWDGYQHNSGEVYIGWPLISPPFSKVVTCQG
QSNKNCSAGNSLLSQLNVLANHLQYFVLMGICGI
>YAPH0S1-31032g LIP2d
MNFFKIFLLSAFAAALSIRQFKSTETYHIDQNAYDFFLKFARLTNIAYCVGPLTAIFKPF
TCGLQCAHFPNMELIRTFRDPAIVFDASGYLAVDHGAKEIYLVTRGTHSIEDVLTDFQVI
QAPLTNMDLAVNITAGDTCKDCRVHSGFLRAYNYTYNQISDTLNSVIKKYPDYKMVVAGH
SFGGATALLFAVNLKVNGHDPLLVTLAQPIVGNRAFANWVDRLFFGQENPDVSKFSPDRN
LYRVTHKGDLVAQVPFWAGYQHSSGEVFIDWPLVHPPLSKVVMCQGQSNSKCSAGNRLRD
QIHLVSSHLQYFVLQGICGL
>YAPH0S5-13278g LIP2e
MKLTNLLAVANLFSAAMALPSPFTSRETEVLEKRVYTSTETDPVTQNTYNLLEKFSRIAS
ISYCVRPPLKIMKPFKCGIPCEYFPDVELLQQFHEPEPFVFDVSGYVAVDHGSKQIYVVA
RGTKGLEDTITDLRVMQAPFDHFDLGTNNASVTADCKDCLVHKGFIQAYNNTMKQIQPVL
DSAVNKYPDYKVAVTGHSLGGATALLLGINLKVNGYDPLVITYGQPLVGNAGFANWVDKL
FFGQENPDTTQITADRKFFRVTHKGDIVTQVPFWDGYTHNSGEVFIDWHTLNPPLHKVVA
CKGQSNKNCSAGNKLVQQVYLKNHVLYFIYQGKCNAFNTFSD
>YAPH0S6-00584g LIP2a
MKFLSLAVLAAATLTAALPSPISSSEAAVLERRVYTSTETSQISQDSYNFFEKYSRLANI
AYCVGPLSNINKPFTCGLQCAKFPDVELIQEFRDPALIFDVSGYLAVDHGSKQIYLVVRG
THSLEDVITDLRIKQAPFTPFDMAANISSTDTCDNCLVHNGFIESYNNTYKQIGSKVDAV
IKQYPDYEIVVSGHSLGGAAALLFGINLKVNGHDPLVVTYGQPIVGNAGFANWVDRLFFG
QENPDVSQVTPERRLYRVTHRGDIVPQIPFWDGYQHCSGEVFIDWPLVNPPLSKVVTCQG
QSNKKCSAGNTLLQQLNVLANHLQYFVLQGICGI
>YAPH0S6-03488g LIP2b
MKLFTLALATLATLAAAVPSPITPSEAAVLEKRVYTSTETAEVSEDSYYFLEKYSRLANI
GYCVGPLANIYKPFKCGLQCGKFPSAELIQEFHDPDYVFDVSGYLAVDHGSKEIYLVFRG
THSLEDVITDLRISQAPLTHFDNAANISSHATCDNCLVHNGFIETYNNTHKQVRAQLDAV
VKKYPDYKIAVTGHSLGGAAALLFGISLKLNGLDPLVITYGQPIVGNKAFANWVDKLFFG
QENPDVSKITPERKLYRVTHRGDIVTQIPFWDGYQHNSGEVYIGWPLISPPFSKVVTCQG
QSNKNCSAGNSLLSQLNVLANHLQYFVLMGICGI
>YAAL0S04-06656g LIP18b
MKTFALLGLLASSVATAAATNASYADHYADVHFSNALGEVRPAPPPAAIIDYARGPQDTA
PITADFYNELEYYMNFSQVAACVGYAGIKKPFNCKVEGCKILQGHVELVSMLSPPELQQF
NVSGYIALDHLKKNIVLVTRGTLDIRDILTDVNILYRKMSPVDFAPNLQSENLKCDGCKI
HDGFLHSYLTTKEFVDPIVAQLKSQHPEYGMVVTGHSLGGATALLFGLDYKINGYDPLVI
TSGQPYVGNQAFADYHDTVWFGKPNPDTVTVGGDRKFYRLTHLHDIVPRVPYWGGYMQVS
GEIFIDYHKTHPPVDKVLACAGQLNNHCTHGSPWWTLFNLLDHFWYIEPFVPCF
>YAAL0S10-01222g LIP2
MKLSTLVLTACAALTAAIPSPLSPPEAAALQERIYTSTETSHIDQNAYNFFEKYARLANI
AYCVGPGTKIFKPFKCGLQCAHFPNVELIQEFHDPAFIFDVSGYLAVDHTSKQIYLVIRG
THSLEDVITDIRIFQQPLTNFDLAANISATATCSDCLVHKGFISSYNNTYNQIGERLDTV
IKQYPDYQIAVTGHSLGGAAALLFGINLKVNGHDPLVVTLGQPIVGNTGFANWVDRLFFG
QENPDVSKVTTDRKLYRITHRGDIVPQVPFWDGYQHCSGEIFIDWPLIHPPLSNVVACEG
QRNKQCSAGNLLIQQANVIGNHLQYFVTEGVCGI
>YAGA0E38754g LIP2
MKLSTILLTACATLVAALPSPVAPSEAAALQKRVYTSTETSHIDQDSYNFFEKYARLANI

GYCVGPGTRIFKPFNCGLQCAYFPNVELIQEFHDPLLVFDVSGYLAVDHASKQIYLVIRG
THSIEDIITDIRIMQAPLTNFDVAANISSTATCDDCLVHNGFIQAYNNTFNQIGPKLDSV
MTQYPDYQVAVTGHSLGGAAALLFGINLKLNGHDPLVVTLGQPIVGNEGFANWVDKLFFG
QENPDVSKVSKDRKLYRITHRGDIVPQIPFWDGYQHCSGEVFIDWPLIHPPLSSVVMCQG
QSNKQCSAGNTLLQQANVIGGHLQYFVTEGVCGI
>YALI0A20350g LIP2
MKLSTILFTACATLAAALPSPITPSEAAVLQKRVYTSTETSHIDQESYNFFEKYARLANI
GYCVGPGTKIFKPFNCGLQCAHFPNVELIEEFHDPRLIFDVSGYLAVDHASKQIYLVIRG
THSLEDVITDIRIMQAPLTNFDLAANISSTATCDDCLVHNGFIQSYNNTYNQIGPKLDSV
IEQYPDYQIAVTGHSLGGAAALLFGINLKVNGHDPLVVTLGQPIVGNAGFANWVDKLFFG
QENPDVSKVSKDRKLYRITHRGDIVPQVPFWDGYQHCSGEVFIDWPLIHPPLSNVVMCQG
QSNKQCSAGNTLLQQVNVIGNHLQYFVTEGVCGI
>YAYA0S2-08614g LIP4
MPAYDLTLGQVLAYLASMFSGQVDATSSSTRIQATQELYNFTAKFSRLANIAYCVDAPLT
PLRTDFTCGESCRYFPNITLDSVIGGDFYSTSITGYIAYDHAKKEKYLVFRGTFSIPDAV
TDIQFQNAPWLTSLPTNLIPNQNDKQTIGKEYVAENKGLGGLEERQAIVHEDPSLVPNKT
ATCDNCQIHSGFADAFNETLRNAGSNFDRFLRNNTDYTMYVLGHSLGAAQAQLFATRFKL
LGYDPHLINLGQPRVGNAEFAAYINQLWFNDTGLIVNDARRLYRLTHWNDVVVGVPDWLN
YTHSIGEVYIDVENVYPTLDKMVVCEGGENPACHRGTFNLWARIDLLQNHLAYIYYIGYC
ALNIGRRDILNLPKYHGNVSYQHGSDPNYNYDTKAPTKLK
>YAGA0E25950g LIP4
MAGFNFTFGQVMSYLLSMFYGQVDATSSSTRIQATQDLYDFTAKFSRLANIAYCINTPFV
PLRTDFTCGESCRYFPDLQLDSVFGGNFSSASTTGYIAYDHKKKEKYIVFRGTFSIPDIV
TDIQFQTAPWLASLPNHLIPTKEDFEHKQAIMKHYAAENKGLGNLEERQDVVHEDPSLIP
KKMDKCENCAIHDGFAKGFNETMEHAGPQIEKYLGNNTDYKMFVIGHSLGAAQAQLFATQ
FKLLGYDPYMINFGQPRLGNPEFAAYINQLWFNDTGLTVNEARRFYRVTHWNDIVVGVPD
WLNYTHSIGEVFIDEESVYPKLDKVVVCEGGENPLCHRGTFNLWQRINFLQNHLAYIFYI
GLCAFNIGRRDVLNMPTYHGNFSYRNSTDPNYDYNTKAPTRITYPN
>YALI0E08492g LIP4
MAGFNFTFGQVISYLASMLYGQVDATSSSTRIQATQDLYDFTAKFSRLSNIAYCINAPFT
PLRTDFTCGESCRYFPDLQLDSVFGGNFSSASTTGYIAYDHKKKEKYIVFRGTFSIPDII
TDIQFQTAPWLTSLPTHLIPTKEDFEHKQAILKHYAAENKGLSNLEERQDVVHEDPSLVP
KKMDKCENCQIHDGFAKGFNETIEHAGPQIEKFLGNNTDYKMFVVGHSLGAAQAQLFATQ
FKLLGFDPYMINFGQPRLGNPEFAAYINQLWFNDTGLVVNDARRFYRVTHWNDIVVGVPD
WLNYTHSIGEVFIDEESVYPKLDKVVVCEGGENPLCHRGTFNLWSRINFLQNHLAYIFYI
GLCAFNIGRRDVLNMPQYQGNFSYQHNIDPNYNYDTKVPTRISKSN
>YAYA0S2-02982g LIP5
MKFSATLLLCAGTFASAAAIGPLLRRDNQTQIINGFEITDFFNDGEEFLYIKDLTNSSSL
GDNNDLNSRDILGAWESSDTAAVSPADFDYLERQAKLANIAYCGGSSLLTIPFSCAYQCK
EFPNMTLVTTWGDSDTVSPLVAGYLSIDHNAKEIVVGFRGSHTLKDWIVDLIVIREAVDT
AYPGCDDCRVHMGFYDSYQATLAEFETDLKTLAAENPGYRLNVVGHSLGGAVALLAATEF
KRQGYNTYLTTFGQPVVGNTAFAKHVNNHWFGSETPNTLEGNSSRQYYRVTHDSDIVPRV
PFWPGYTPNAGEVYIDVPQIGPVVSDLQFCDGEINYQCVYGNSLLSLISLTAHHNYFVYI
GGDC
>YAYA0S4-04016g LIP5p [pseudogene]
MKF*TTILLCTGTLVASAAIDPVVALDNQTEIFHDYEILDNQTEIINGFEITDFFNEEGE
EFIFIKDLTNSSLLEDNFEDILSSRKLGAWKSSTTGAVSQTSFDYLQKQAKLANIAYCGG
ASLMSAPFICAYQCKEFPNMTLASTWGNSDNHSPLVAGYLAVDDSEKSIVVGFRGSHTLK
DWILDLLVIRRAVDHAYPGCSGCRVHQGFYSAYVATLAYFDADLKKLVAENPGYRVNVVG
HSLGGAVAVLMAATFQNRGYDTYLATFGQPVVGNTAFANYLDRLWFGSENPKTLGDDPSR
RYYRVTHKSDIVPRVPFWPGYTPSSGEVYINVPQTDPSVSDLKTCDGQLNYKCIYGNSLL
SLISMTAHSNYFVQMGGDC
>YAAL0S02-16094g LIP5
MKLSTALLSLCLLGGVSAAPTLGDATKISVVNGFQVAEFIGEDGTEKIVFTDLLNSTWAA
PDLSGLTRSLLGYPEDSGDAEDSDLEERAPVAKALTQETYNYLEKQVKLANIAYCGGTSQ
LISPFVCAYQCKEFPTVELVTTWDTDGLNVSPAVAGYLAIDHESKEFILGFRGSKTLKDW
LVNLNTIRVPVNKKYAPCKGCEVHLGFYNAYKATLGIFESTLTYLRSKHPEYRLNVVGHS

LGGAVALLVATDFKQRGYDTQLTTFGQPIVGNTKFANHIDNLWFGTQAESSTGQFHRVTH
RNDIVPLVPFWLGYSQTAGEIFIGAPNLAPPLDSLRVCEGRNSRECLNGNSILGLLELSA
HTEYFVALGGDC
>YAGA0E32462g LIP5a
MKFSATFLLCTVSLVAAATIAPLAPRDNQTKVINGFQISDLVNDDGDEFLLFKDLTNSSW
PGDTEEDNLYSGVIKAWESTEVAAVSQTSYSYLDKQAKLANIAYCGGSSLLTSPFTCAYQ
CKEFPNMKLVTTWGDSATLSALVAGYLSIDHTDKEIVVGFRGSHTLKDWIVDLMVLRKAV
DDSYPGCDDCRVHHGFYNAYKATLPKFDNDLKKLVAENPGYRVNVVGHSLGGSVALLAAT
EFKNRGYDTYLTTFGQPVTGNTAFANYVDNLWFGSETPKTLSGDSSRRYYRVTHKSDVVP
RVPFWAGYTQNAGEVYIGVAQIDPPVSSLHYCEGEVNHQCVYGNSLLSLISLTAHHNYFV
YIGGDC
>YAGA0F11166g LIP5b
MKLLTTTLIITLASAATLPLISTNSTTIINGFQISNLINDDGDELLLFEDLTKDNVFANS
SLKSDSTLSIRQSSTSPISQTTYNHLVRQSKLANIAYCSGPPSLIAPFTCSYQCKEFPNM
NLVSTWGEDPLSPSVSVAGYLSVDHTEKTIVVGFRGSQTFKNWIVNFMVLRNPVENSYQG
CEGCTVHQGFYNAYKDTKKEYDTELIKLVEENPGYIVNVVGHSLGGSVALLAATDFKNRG
YKTSLTTFGQPVTGNTAFANYIDNLWLTNDTKNTTRSYYRVTHKGDVVPRVPFWHGYSPT
AGEVYIGAAELNPPVSSLLDCEGQENEQCIQGNSLLSLVNLTAHNLYFVYLGGECEVTSL
>YALI0E02640g LIP5
MKFSATLLFCASTLIAAATIAPLAPRGNQTKIINGFQVTDLVNDAGDELLLFKDLTNSSW
PGDTSTSTNGVTLYSGAIKAWTSSDVAAVSQTSYNYLEKQAKLANIAYCGGSSLLTAPFT
CAYQCKEFPNMKLVSTWGDSRTLSALVAGYLSIDHTDKEIVVGFRGSHTLKDWIVDLMVL
RKAVDDSYPGCDNCRVHHGFYSAYKATLARFDNDLKKLVAENPGYRVSVVGHSLGGAVAL
LAATDFKNRGYDTYLTTFGQPVVGNTGFANYVDDLWFGSETPNTLSGDSSRRYYRVTHKS
DVVPRVPFWPGYTPNAGEVYISVAEINPPVSALRYCDGEVNDQCLHGNSLLSLINLTAHH
NYFIYIGGDC
>YAAL0S04-02696g LIP7
MVQIGKFTEWLSVTLWGAAATTSSTATSSITQNTYDFVRTFSHLSNVAYCVKAPIKSLDD
NFQCGNACKNFPNMELVTTFGGDFFQTSITGFLALDHVKKEKYVVYRGTFSIADVITDLQ
FQQSGFLVDAPALNSLKANDTSESAKIDCKDCKIHDGFKKANTETMTNIGDDLKKHLDSY
PDYKLYVTGHSLGAAQALLSAISIKLQGYDPTLINFGQPRVGNAAFANYVDRLFFGEDAG
LSVTSDRKLYRLTHWNDVFVGLPNWDGYQHNVGEVFIDWRFTNPPLQYVKSCEGGENPKC
YRKDFNLLAQINLLQNHLAYIDYIGYCTLNIGRRAQMNLPRYTGPNTYAHKTEDDVELAL
Y
>YAGA0D10484g LIP7
MVNFGARVKDFFSVLLFGAASTTSSTKTALVSQGFYDAALDFSHLSNIAYCVNAPITPLK
DDFSCGQSCVHFPDIELVHTFGGDFFSTSITGYLALDHVKKEKYVVFRGTFSIADAITDI
QFQQSSFLVNVPALNTFTANDTSAEAQIDCKHCKIHDGFSKAFTETLHNIGPQLKQHLDS
YPDYQLYVTGHSLGAAMALLAGTSIKLKGYDPILINYGQPRVGNKAFADYISTLWFGKGD
GLEINKDRRLYRMTHWNDVFVGLPNWDGYTHSNGEVYIKGKYVNPPLKDLMSCAGGENSK
CYRSTFNLLAQINLLQNHLCYIDYIGFCALNVGRRQLNDMPHYTGPYTYGHKTEEDFVAE
GLELSN
>YALI0D19184g LIP7
MVSFGARIKDFFSVLLFGAASTSSSTKTALVSQGFYDAALDFSHLSNIAYCVNAPITPLK
SDFSCGQSCVHFPDIELVHIFGGDFFSTSITGYLALDHVKKEKYVVFRGTFSIADAITDI
QFQQSSFLVNVPALNTFIANDTAPEAQIDCKQCKIHDGFSKAFTETWHNIGDLLEQHLDS
YPDYQLYVTGHSLGAAMALLGATSIKLRGYDPILINYGQPRVGNKAFADYISALWFGNGD
GLEINQQRRLYRMTHWNDVFVGLPNWDGYTHSNGEVYIKGKYVNPPLKDVFSCAGGENSK
CYRSEFNLLAQINLLQNHLCYIDYIGFCALNVGRRELNDLPHYNGPYKYGHKTEEQFIAE
GLELSN
>YAGA0F22892g LIP8
MVSLSARIKDFFSVLLLGAAEITPSTQTAGVSQGFYDFARDFAHLSNVAYCVDAPITPLK
ADFTCGNACKHFPDIELVKTFGGDFFDTSITGFLAVDHVKKEKYVVFRGTFSLADAITDM
QFQQSPFLVDVPAMNTFSANDTAAEAQTQCEGCKIHDGFSKAFTETWGEIGEDLHKHLDS
NPDYQLFVTGHSLGAAMALLGATSIKLKGYDPILINYGQPRVGNKPFSEFINKLWFGDDN
GLEIKPERRLYRMTHWNDIFVGLPNWEGYTHSNGEVYINNRFINPPLKDVISCAGGENSQ
CYRSSFNLLSQINLLQNHLAYIDYIGYCALNIGRRELADQPHYKGNYFYAHRTEEDFKKL

GLELSTRTKQ
>YALI0B09361g LIP8
MVSLSARIKDFFSVLLLGAATITPSTQTAGVSQGFYDFARDFAHLSNIAYCVNAPITPLN
PDFTCGNSCKHFPEIELVKTFGGNFFKTSITGYLAVDHVKKEKYVVFRGTFSLADAITDM
QFQLSPFLVDVPALNTFSANDTTAEAQTHCEGCKIHDGFSKAFTETWGNIGEDLQKHLDA
NPDYQLYVTGHSLGAAMALLGATSIKLKGYDPILINYGQPRVGNKPFAEFINKLWFGEGN
GLEITPERKLYRMTHWNDIFVGLPNWEGYTHSNGEVYINNRFINPPLKDVISCAGGENSK
CYRSSFSLLSQINLLQNHLAYIDYIGYCALNIGRRELADQEHYTGPYYYGHRSEEDFKKL
GLELSTPQVEN
>OLHI0E00452g LIP9
MLIWLPLVLCVSQKLLAELTRYMRFAEVAGCVGSGSRGQKLPGITRPFQCPLCDLPVFGS
AELLGTFSSAYFEKQPTGYIAVDTYHGELLVVFRGTTSFGDKIADFSIFKKRVCEVEALQ
EHLDMIDDNAQIFGGVVDVFNGGFEPLIWQIKRALMKYPGYRLVVLGHSLGGAIATLCAV
TLKLRGFNPYLITFAEHLTGNWEWAQTVNWLFPGKTSYRVTRVGDIVPRLPPTELGYYPR
GSEYYISSYEMPLHHEELFYCDESNHCSQRAPLWSYLLNYQETQKSHFNYFFELRKCTA
>YAYA0S2-34200g LIP9p [pseudogene]
MRLATIASFVGLVTASPIGWIPPALLNPFGPRKPDGTVVNATLATVNEMEHYWKFCSVSY
CVGIGNNNQLYSQIKTPFVCNNILCADEEFSETELLYSFYGVNEHQTANGYIAVDHKAKQ
LVLVFRGTQDEADTAADLNTWQVSNVDFSGITNASDTNAQSSCPGCSLHAGFVGIFNYSF
KQINGRLNLYKATYPDYKLVVTGHSLGGAIALLYGVSLKINGRDPLVVTFGQPRVGNAAF
ATYVDSLFLPTAGDQLTTSPSRNMYRVTRYEDPVTQVPFWAGYTQQSGEVYINQFSVPTN
AENVMLCQGQNNGFCANGIPWYQYANLDSDKQLHSSYFFRSPSCGGSQSYVPYGSNNTEP
TSAITIPPGSLQNQLSLPYDIPN
>YAAL0S08-03510g LIP9
MRIAIFASLLALVQASPIGIVPPVFLPLLGGKGADGTIAKAPVSIVDEMEHYWKFCSVSY
CVGMGKNDQLYSQVRQPFVCNNVLCADQEFSQTEVLYSFYGINEHQTANGYIAVDHKRKQ
LVLVFRGTQSEADSAADLNTWQVPNVNFDGLKNTTGTNAESDCQGCSIHAGFVGIFNNSF
KAIDSRLNLYKAQYPDYKLVVTGHSLGGAVALLYGISLRVNGRDPLVVTFGQPRVGNEAF
ANYVDSLFFPSLNDMLSYSPYRKMYRVTRYEDPVTQVPFWDGYTQQSGEVFINQFKVPTD
PENVLFCQGQNNRNCANGIPWYQYANIASDKSVHSSYFFRSPSCSNAQAYTPYGGNTTAA
PGAVSLDPSELKTRLNLPYDF
>YAGA0E00738g LIP9
MKLAAIASFVGLVAASPIGLLPPALLNPFGPKKPDGTVAPAPLDVVTEMEHYWKFCSVSY
CVGMGKNNQLYSQVKTPFVCNNILCADKEFSETELLYSFYGINEHQTANGYLAADHKRKQ
LVLVFRGTQSEADSAADLNTWQVSNVDFDGLKNSTSTNAESECQGCSIHAGFVGIFNNSF
KQIDSRLNLYKGMYPDYKLVVTGHSLGGAVALLYGVSLKINGRDPLVVTFGQPRVGNAAF
ASYVDSLFFPTASDQLSSSPYRKMYRVTRYEDPVTQVPFWDGYTQQSGEVFINQFKVPTD
PKNVVFCQGQNNGYCANGIPWYQYANIDSDKQVHSSYFFRSPGCAGTQSFTPYGGNTSEP
SAVSLTPGDLKARLNLPYDTNY
>YALI0E34507g LIP9
MYNGIAWEGYIYILGPFLKTRQHQKPQDTIMRLATIASFVGLVTASPIGLLPPALLNPFG
PKKPDGTVATAPLDVVHEMEHYWKYCSVSYCVGMGKNNQLYSQVKSPFVCNNILCADQEF
SQTELLYSFYGINEHQTANGYLAADHKRKQLILVFRGTQSEADSAADLNTWQVSNVDFDG
LKNSTDTNAESDCHGCSIHAGFVGIFNNSFKQIDSRLNLYKSMYPDYKLVVTGHSLGGAV
ALLYGVSLRINGRDPLVVTFGQPRVGNAAFASYVDSLFFPTAGDQLSSSPYRKMYRVTRY
EDPVTQVPFWDGYTQQSGEVFINQFNVPTKPENVVFCQGQNNGFCANGIPWYQYANIDSD
KQVHSSYFFRSPGCGGSQSFTPYGGNQTEPSAVSIPPGDLKARLNLPYDTNY

V. CONCLUSION

Dans les chapitres précédents, nous avons reconstitué le métabolisme lipidique à partir des 6 génomes disponibles, du clade et de OLHI. Nous avons pu confirmer l'importance de ce métabolisme pour ces organismes oléagineux en constatant que les protéines impliquées dans le métabolisme lipidique sont plus conservées et ont une pression de sélection purifiante plus élevée que le reste des génomes. Nous avons ensuite voulu savoir s'il existe au sein du métabolisme lipidique des points chauds pour l'évolution qui pourraient faire ressortir une voie particulière. Mais les gènes ayant des taux de conservation ou les pressions de sélections les plus forts ou les plus faibles ne constituent pas une voie métabolique particulière et sont répartis dans tout le métabolisme lipidique. Ce résultat semble toutefois biaisé par la présence des gènes en famille mais si l'on retire ces gènes de l'analyse nous n'avons plus assez de données pour répondre à la question puisqu'ils représentent environ 65% des gènes étudiés.

La reconstitution du métabolisme lipidique a, en effet, révélé un nombre conséquent de grandes familles multigéniques impliquées dans ce métabolisme suggérant que les pressions agissant sur ces protéines sont différentes et permettent des duplications qui peuvent être suivies de sub- ou néo-fonctionnalisation ou de pseudogénisations. Dans ce chapitre, nous avons alors voulu étudier plus particulièrement le dynamisme et l'évolution des familles de gènes. Pour étudier l'évolution des familles de gènes, nous avons identifié les orthologues dans chaque espèce du clade et OLHI et nous avons reconstruit les scénarios évolutifs afin de regrouper ensemble les gènes descendants d'un même gène ancestral. Nous avons alors classé les familles en fonction de leur dynamisme et ainsi repéré 8 familles très dynamiques. Nous avons ensuite déterminé la conservation et les pressions de sélection s'exerçant sur les protéines en famille.

Expansion ou contraction des familles de gènes ?

Les reconstitutions des scénarios évolutifs des familles nous ont permis de mettre en évidence que l'expansion de ces familles de gènes a principalement eu lieu après la divergence de YAHI et avant l'apparition de l'ancêtre commun au clade, qui possède déjà 76 gènes en familles dynamiques. Cette expansion semble continuer soit de façon espèce spécifique soit par groupe d'espèces notamment pour YALI-YAGA dont l'ancêtre commun a subi 4 duplications. Par contre chez YAPH nous observons un phénomène de contraction des familles qui apparaît comme un

retour en arrière avec 23 pertes spécifiques. De même, YAYA a subi 11 cas de pertes spécifiques et compte un plus grand nombre de pseudogènes suggérant que ce génome va aussi vers une contraction des familles comme YAPH. Cette étude était limitée aux gènes impliqués dans le métabolisme lipidique, mais il serait intéressant de voir si nous observons les mêmes tendances pour les autres familles multigénétiques des génomes. Il serait très intéressant de ne pas observer de phénomène d'expansion des familles non lipidiques, notamment chez les espèces pour lesquelles les familles lipidiques sont encore en expansion, car ceci révèlerait une véritable adaptation à leurs milieux de prédilection riches en lipides.

La famille des lipases

La famille des lipases, avec ses 15 duplications, 11 pertes et 4 pseudogénisations qui ont eu lieu au cours de l'évolution de ces 10 lipases ancestrales est la famille la plus dynamique de notre étude. De plus, les lipases sont des enzymes très intéressantes pour des applications biotechnologiques diverses. Nous nous sommes donc plus particulièrement intéressés à cette famille. Nous avons ainsi pu mettre en évidence l'existence d'une nouvelle lipase n'existant que chez YAAL qu'il serait très intéressant de caractériser fonctionnellement dans le cas où elle serait spécifique à l'écologie de YAAL.

Le plus remarquable dans l'étude de cette famille est la grande expansion du nombre de gènes notamment chez YALI et YAGA et plus particulièrement pour les LIP2 de YAPH. Cette observation nous a donc amené à nous poser les questions des fonctions de chacune de ces lipases et de leur régulation. Mais de façon surprenante, seulement 19 des 39 lipases, pour lesquelles nous disposions de données de transcriptomique, sont exprimées dans au moins une condition. De plus, les lipases descendant directement de lipases ancestrales au clade sont les seules exprimées et également les mieux conservées. Nous avons pu observer que la conservation des promoteurs entre deux lipases d'un même groupe monophylétique n'est que très rarement visible. Ceci suggère que soit seul les transcrits sont dupliqués, soit les promoteurs aussi sont dupliqués mais ils évoluent de façon neutre et ne sont donc pas reconnaissables.

Toutes ces constatations amènent à se poser la question de l'intérêt de dupliquer ces lipases. Nous pouvons penser que nous n'avons pas trouvé les conditions dans lesquelles elles s'expriment, ce qui est tout à fait possible puisque nous n'en avons testé que 3 différentes. Il serait intéressant, pour chacune des lipases du clade, de tester une multitude de conditions. Pour cela il faudrait insérer individuellement chaque gène dans une souche délétée pour toutes les autres lipases. Si les outils génétiques développés pour YALI sont transférables sur OLHI, il serait alors beaucoup plus simple d'utiliser OLHI pour cette expérimentation puisqu'il n'y aurait que 3 lipases à supprimer. La croissance des mutants ainsi obtenus pourrait être testée sur une multitude de substrats et si la souche croît cela signifie que la lipase insérée s'exprime.

Des dN/dS surprenants

Nous avons déterminé les pressions de sélection qui s'exercent sur tous les orthologues synténiques et sur tous les gènes du métabolisme lipidiques dont nous savons qu'ils descendent d'un même gène ancestral. Pour cela nous avons calculé les dN/dS pour 14315 couples de gènes et nous n'avons trouvé qu'un seul gène ayant un dN/dS supérieur à 1. Traditionnellement, un ratio supérieur à 1 implique une sélection positive, un ratio inférieur à 1 implique une sélection purifiante, et un ratio égal à 1 indique qu'il n'y a pas de sélection. Les valeurs très basses observées pour le clade *Yarrowia* pourraient s'expliquer de plusieurs façons. La première explication serait que les gènes de ces génomes aient subi beaucoup plus de périodes de sélection purifiante que de sélection positive ou que les différents sites des séquences de ces gènes sous majoritairement soumis à une sélection purifiante plutôt qu'à une sélection positive. Ceci signifierait que les gènes de ces génomes sont globalement sous une forte pression de sélection purifiante. Une autre explication serait que ces génomes ont des caractéristiques propres qui impliquent que la méthodologie de calcul du dN/dS n'est pas adaptée. Ceci nécessiterait de changer la limite établie indiquant une sélection neutre, c'est à dire dN/dS=1, et séparant la sélection purifiante de la sélection positive.

Dans une étude des pressions de sélection chez deux hémiascomycètes du des clade CTG, *C. albicans* et *C. dubliniensis*, les dN/dS obtenus par les auteurs sont dans plus de 90% des cas inferieurs à 0,3 et seulement un gène, qui serait impliqué dans la formation de biofilm, a un dN/dS supérieur à 1 (3.077) (Jackson et al., 2009). De même dans une étude sur les génomes mitochondriaux de 8 espèces du clade *Lachancea*, les dN/dS calculés pour 8 gènes sont toujours inférieurs à 0,8 (Friedrich et al., 2012). Ces deux études menées sur différents clades de levures concluent à une forte pression purifiante agissant sur ces génomes et sont cohérentes avec nos résultats. Mais puisque nous observons tout de même des pressions purifiantes très forte, avec un dN/dS moyen de 0,07, sur tous les gènes en familles multigéniques dynamiques, nous pouvons nous demander si ces gènes peuvent vraiment être tous sous ce type de force alors qu'ils appartiennent à des familles qui ont subi beaucoup d'évènements et qu'ils ont dans la plupart des cas développé des spécialisations ou des nouvelles fonctions depuis leur divergence avec le gène ancestral. Dans le cadre d'une famille dynamique, une très forte pression purifiante sur les différentes copies serait cohérente si les duplicats servent de répétition de dose et donc gardent la même fonction. Une autre possibilité serait que l'évolution de ces gènes en familles soit très ancienne et que leur séquence soit, aujourd'hui, effectivement figée par une sélection purifiante. Pour tenter de répondre à cette problématique nous pourrions calculer les dN/dS de tous les gènes provenant d'une duplication d'un gène chez l'ancêtre commun à deux espèces très proche dans le clade. Si nous constatons qu'ils sont sous une sélection purifiante relaxée nous pourrions conclure que les dN/dS très bas obtenus dans cette étude révèlent bien une fixation des séquences des gènes en famille.

DISCUSSION ET PERSPECTIVES

YALI un modèle compétent ou historique ?

Les organismes oléagineux sont étudiés depuis très longtemps pour la capacité à stocker et à synthétiser de larges quantités de lipides, par définition, plus de 20% de leur poids sec. Un des principaux enjeux de la recherche sur les organismes oléagineux est la production de biocarburants. Cet enjeu est crucial, puisque cela permettrait de s'affranchir des aléas de la filière pétrole, tout en contribuant à la limitation des émissions de gaz à effet de serre puisque les biocarburants génèrent beaucoup moins de gaz à effet de serre que les carburants actuels. Chez les plantes, beaucoup d'organismes sont reconnus oléagineux ou à fruits oléagineux comme par exemple le colza, le tournesol, l'olivier, le soja et le lin. Parmi les bactéries nous pouvons citer *Rhodococcus opacus, les* actinobactéries du genre *Mycobacterium*, *Streptomyces*, *Rhodococcus* et *Nocardia* et les cyanobactéries (Alvarez and Steinbüchel, 2002), Certaines micro-algues diatomées et chlorophycées sont aussi des organismes modèles pour ces études. Des champignons filamenteux sont aussi capable d'accumuler de grandes quantités de lipides, comme *Mucor circinelloides* et *Mortierella alpina*. Enfin, parmi les levures nous pouvons citer plusieurs espèces appartenant aux ascomycètes et basidiomycètes : *Apiotrichum curvatum*, *Candida freyschussii*, *Rhodotorula glutinis*, *Rhodosporidium toruloides, Cryptococcus curvatus*, *Lipomyces starkeyi* (Ageitos et al., 2011; Huang et al., 2014) et *Yarrowia lipolytica*. Même si le niveau d'accumulation des lipides de YALI est inférieur à celui d'autres organismes cités précédemment, elle est devenue un organisme modèle car elle est facilement cultivable et manipulable génétiquement. Les outils développés depuis des dizaines d'années sur cette levure sont stables et permettent d'envisager tout type de clonage et d'expression de gènes cibles. Les différents types d'approches utilisés au cours de cette thèse ont toutefois mis en évidence le potentiel de OLHI comme modèle pour la production de lipide. Nous avons en effet constaté qu'elle a de meilleures capacités de croissance sur des substrats hydrophobes et notamment sur acide oléique, et stocke plus de lipides que YALI. Le génome de OLHI est aussi beaucoup plus petit ce qui diminue les coûts de séquençage et facilite les manipulations. De plus elle possède les mêmes voies impliquées dans le métabolisme lipidique que YALI et toutes les mêmes fonctions sont assurées par ses 160 gènes lipidiques au lieu des 204 de YALI. En effet, l'expansion des familles de gènes impliquées dans ce métabolisme s'est produite après la divergence de OLHI ce qui diminue la redondance de ses gènes lipidiques à 1,5 contre 1,9 pour YALI. Chez YALI, beaucoup de gènes appartenant a une famille nombreuse ne semblent pas être

exprimés dans les 3 conditions testées durant cette thèse, c'est-à-dire glucose, acide oléique et tributyrine. Donc soit ces gènes ne sont jamais exprimés soit ils sont exprimés dans des conditions particulières. Dans les deux cas la délétion de tous les gènes d'une famille reste nécessaire dans le cadre d'étude de la spécificité et de la fonction d'un des gènes appartenant à la famille en question. Pour l'étude des lipases par exemple, il suffirait de déléter 3 LIP chez OLHI contre 16 chez YALI ! D'autre part, la recombinaison homologue chez YALI est très peu efficace, nous pouvons alors penser qu'elle pourrait être meilleure chez OLHI. Dans ce cadre, OLHI s'impose comme un modèle qui mérite d'être testé pour savoir si les outils développés sur YALI sont transférables sur OLHI. Si OLHI n'est pas facilement manipulable génétiquement, elle reste tout de même un réservoir de gènes à tester dans YALI car avec ses 1300 gènes de moins dans son génome elle n'en reste pas moins efficace.

Quid de l'émergence du caractère oléagineux chez les levures

La présence de tous les gènes du métabolisme lipidique chez OLHI et la similitude des caractéristiques physiologiques entre OLHI et YALI montrent que le caractère oléagineux est apparu avant OLHI. Ceci soulève alors une question sur laquelle il serait très intéressant de se pencher : peut-on dater l'apparition des capacités oléagineuses ? Si oui, à quand remonte-t-elle ? Il est en théorie possible de remonter dans l'arbre du vivant pour retrouver l'apparition d'un caractère. Mais, comme nous avons vu précédemment, le caractère oléagineux n'est pas monophylétique. Nous savons qu'il existe des gènes spécifiques aux organismes oléagineux qui sont par exemple les deux gènes ACL codant l'ATP-citrate lyase, les 4 ACD codant des acyl-CoA déshydrogénases et les gènes impliqués dans la dégradation de la leucine (Vorapreeda et al., 2012). Ces gènes ont probablement été acquis par transfert horizontal chez un ancêtre des hémiascomycètes et indépendamment perdus dans des ancêtres plus récents. Il faudrait donc tout d'abord remonter de proche en proche dans l'arbre des hémiascomycètes jusqu'à déterminer dans quelles espèces ces gènes sont tous présents et quelles espèces sont oléagineuses. Il faudrait également construire les phylogénies de chacun des gènes afin de déterminer leur présence/absence dans les génomes d'hémiascomycètes séquencés. La réconciliation de ces arbres avec l'arbre des espèces permettrait de reconstruire un scénario évolutif pour chaque gène. La grande limite de cette étude est le faible nombre d'espèces hémiascomycètes connues comparé au nombre réel estimé et surtout le faible nombre de génomes séquencés. Le petit nombre de génomes actuellement disponibles peut être comblé en partie par la reconstruction des génomes ancestraux. Cette étude permettrait de repérer à partir de quel ancêtre le set de gènes spécifiques aux oléagineux est entièrement présent dans le génome et aussi de dater les pertes de gènes conduisant à la perte du caractère oléagineux.

A une échelle plus large, l'accès à une centaine d'ancêtres communs répartis dans tout l'arbre du vivant, qui sont actuellement en cours de reconstruction dans le cadre du projet Ancestrome au LBBE à Lyon, permettrait de faire l'étude de l'apparition du caractère oléagineux chez les plantes, les bactéries, mais probablement pas chez les micro-algues car trop peu de génomes sont disponibles. Ceci pourrait permettre de dater les transferts horizontaux des gènes spécifiques au caractère oléagineux et de déduire le caractère oléagineux des ancêtres. Ce métabolisme ancestral, déduit sur la base de caractères génomiques anciens pourraient ensuite être relié à des évènements ayant affecté l'évolution de la Terre, en corrélant les variables génomiques et une variable environnementale obtenue à partir d'estimations géologiques, par exemple une grande crise biologique suivie d'une longue période de carence des ressources alimentaires amenant à favoriser les nouvelles espèces capables de stocker des réserves nutritives.

L'expansion des familles : utile ou invasif ?

La reconstitution des scénarios évolutifs des familles de gènes impliqués dans le métabolisme lipidique nous a permis de conclure que ces familles sont probablement encore en train de s'étendre chez YALI, YAGA et YAAL. *A contrario*, les familles de YAPH et YAYA tendent plutôt vers une contraction des familles et ceux ci possèdent entre 15 et 20 gènes lipidiques en famille de moins que YALI. Pourtant ces espèces semblent toutes aussi performantes que YALI pour stocker et synthétiser des acides gras et toutes les fonctions pour le métabolisme lipidique sont assurées par ce plus petit nombre de gènes. OLHI, qui n'a pas subi une expansion de ces familles est même plus performante que YALI. Ceci nous amène donc à nous demander quel est l'intérêt de dupliquer ces gènes lipidiques ? Pourquoi ces familles possèdent-elles autant de paralogues ? Dans les espèces YALI, YAGA et même YAPH nous avons pu observer que parmi ces gènes dupliqués souvent seule la copie ancestrale s'exprime dans les milieux testés et dans la plupart des cas seuls les transcrits semblent s'être dupliqués. Nous pourrions donc penser que ces copies se dupliquent de façon invasive dans le génome et sont des séquences non utilisées. Pourtant les grandes familles correspondent à des gènes codant des étapes clés du métabolisme lipidique, telles que les lipases ou les acyl-CoA oxydases. De plus, même si ces familles sont dynamiques les pressions de sélection qui s'exercent sur leurs gènes sont plus purifiantes que celles agissant sur le reste des génomes et les séquences de ces protéines sont plus fortement conservées. Nous n'observons d'ailleurs que 12 cas de pseudogénisations de gènes lipidiques pour l'ensemble des génomes traités, suggérant que des pressions agissent bien pour la conservation de gènes dupliqués.

Une explication à ces expansions de familles pourrait résider dans la grande variabilité des

milieux dans lesquels ces espèces sont capables de vivre. En effet, les géographies et les milieux d'isolements des souches du clade *Yarrowia* actuellement décrites sont très diversifiés, allant de l'estomac de termite pour la souche CBS 10253 de YAYA à de l'eau de mer pour la souche CBS 10407 de YAPH. Chacun des isolats d'une même espèce ont également été découverts dans des environnements complètement différents. Par exemple, les deux souches de *Yarrowia porcina*, récemment décrites, ont été isolées dans de la viande de porc dans un cas et dans des sédiments de rivière d'eau salée tropicale dans le second cas. Ceci pourrait suggérer que ces espèces possèdent une diversité de gènes paralogues dont chacun s'exprime dans une condition très spécifique. Dans ce cas, leur expression est donc très difficile à observer en laboratoire car il faudrait tester un très grand nombre de substrats avant d'assigner l'expression d'un gène en famille à une condition particulière. Cette grande diversité implique qu'il est très difficile de faire des liens entre la physiologie des espèces et leur contenu en gènes. Cependant, l'accroissement des données de métatranscriptomiques pourrait permettre d'accéder bientôt à des transcriptomes séquencés d'espèces du clade *Yarrowia* ayant poussé dans leur milieu naturel. Si ces copies de gènes ont bien une fonction et ne sont pas des séquences « inutiles » nous devrions alors observer, dans ces conditions naturelles, l'expression d'au moins une partie de ces gènes que nous n'observons pas dans les conditions testées en laboratoire.

Actuellement, le clade *Yarrowia* semble en plein essor avec de plus en plus d'espèces décrites. La découverte de ces nouvelles espèces sera probablement suivie par des séquençages de leur génome. Ces nouveaux génomes auront un grand intérêt pour mieux comprendre la fonction et l'évolution des gènes ou des familles impliqués dans le métabolisme lipidique et dans d'autres métabolismes, comme celui de l'assimilation des sucres. Au delà de ces études du métabolisme, ces nouveaux génomes nous permettrons également de mieux comprendre l'évolution de leur taille et probablement de relier la contraction des familles de gènes de YAPH et la diminution de la taille de son génome. De façon plus générale, la disponibilité d'un plus grand nombre de génomes du clade *Yarrowia* permettrait en effet de déterminer les mécanismes à la base de l'expansion de la taille des génomes de ce clade.

Une perspective : Comprendre l'expansion de la taille des génomes

Le séquençage de quatre espèces du clade *Yarrowia* et de l'espèce out-group OLHI a permis d'observer des différences de taille des génomes allant du simple au double avec 10,7 Mb pour OLHI et 22,8 Mb pour YAGA. Au sein des eucaryotes, les tailles des génomes varient énormément, avec des génomes d'environ 12,5 Mb chez les levures, 13,6 Mb pour SACE (Goffeau et al., 1996), et allant jusqu'à 150 OOO Mb pour le monocotylédone *Paris japonica* (Pellicer et

al.) mais cette distribution est largement biaisée vers les petits génomes. La taille du génome est fortement corrélée à un certain nombre de traits phénotypiques, tels que la taille des cellules (Gregory, 2001), le taux du métabolisme basal (Kozłowski et al., 2003), et au nombre relatif de gènes, d'introns, et d'éléments transposables (Lynch and Conery, 2003b). Ceci suppose que le faible nombre de grands génomes chez les eucaryotes est due à une élimination sélective des séquences dites « inutiles » ou « inutilisées ». Afin de comprendre les mécanismes à l'origine de ces différences de taille des génomes, de nombreuses études ont été menées mais le plus souvent à des niveaux taxonomiques élevés, c'est-à-dire en comparant des organismes éloignés. Les études sur la variation de la taille des génomes entre des espèces proches restent rares et dans la plupart des cas elles sont menées sur des espèces de drosophiles (Bosco et al., 2007; Boulesteix et al., 2006). Pourtant, la variation à ces niveaux taxonomiques inférieurs pourrait fournir des indices importants sur le tempo et le mode d'évolution des génomes. Le clade *Yarrowia* et son out-group seraient donc des modèles de choix pour tenter de répondre à la question de l'évolution de la taille des génomes. Nous avons constaté que la variation de taille dans ce clade est principalement due à des quantités différentes d'ADN dans les régions intergéniques. Dans notre cas, la différence de taille du génome n'apporte pas nécessairement de nouveauté en terme de gènes. Ayant écarté l'activité des éléments transposables, la duplication du génome et le nombre d'introns comme mécanisme aboutissant à un changement de taille nous pouvons proposer plusieurs mécanismes. Le premier serait la création de microsatellites par l'insertion des bases supplémentaires dans le brin d'ADN néosynthétisé par le complexe enzymatique de réplication. Mais ce mécanisme n'expliquerait pas une différence de 10 Mb et un biais de composition en microsatellites dans le génome de YALI n'a pas été mis en évidence lors de son séquençage et de son analyse comparative à d'autres levures (Barth et al., 2013). Le second mécanisme pourrait être le transfert d'ADN mitochondrial au noyau appelé NUMT. Mais les NUMT ont été quantifiés chez YALI et ils ne représentent que 2 kb de son génome (Sacerdot et al., 2008). Le mécanisme le plus probable serait des évènements de cassures doubles brin nécessitant une réparation. Les mécanismes de réparation peuvent être classés en deux catégories : la recombinaison homologue, qui utilise une séquence d'ADN identique à la séquence endommagée comme matrice pour la réparation ; et la recombinaison non homologue ou réparation par jonction des extrémités non homologues (NHEJ). Il s'agit de la réunion de deux fragments d'ADN ne présentant pas ou très peu d'homologie de séquence. Dans ce cas, la rupture et la réunion des régions d'ADN impliquées est à l'origine d'intégrations ou de duplications, excisions, inversions ou déplacements de segments d'ADN et peuvent aboutir à des réarrangements chromosomiques. Dans le cas des levures, il s'agit d'un mécanisme de réparation rapide lorsqu'il est impossible de faire de la recombinaison homologue par exemple pendant certaines phases du cycle cellulaire où il y a absence de séquence homologue dans des cellules haploïdes (Pâques and Haber, 1999). Nous savons par ailleurs que la recombinaison

homologue chez YALI est peu efficace ce qui appuie l'hypothèse de cassures suivies par des réparations non homologues.

Ces recombinaisons non homologues permettent également d'expliquer l'augmentation de la redondance des gènes dans les génomes du clade, ce que nous avons notamment observé pour les gènes du métabolisme lipidique avec les 37 familles multigéniques. Ce mécanisme serait aussi en corrélation avec les nombreux réarrangements chromosomiques observés entre les espèces proches qui sont pour la plupart des réarrangements intrachromosomiques. Ce mécanisme est appelé mésosynténie (Hane et al., 2011). La présence de mésosynténie est une originalité au sein des hémiascomycètes car elle y est actuellement peu démontrée. Nous savons seulement qu'elle existe au sein des CTG qui, pourtant, ne semblent pas subir d'expansion de la taille de leur génome.

Si les recombinaisons sont à la base de l'expansion des génomes dans le clade *Yarrowia*, cela suppose que ces génomes n'auraient pas la capacité d'éliminer les séquences non codantes « inutiles » ou « inutilisées » et seraient donc plus permissifs à l'amplification des séquences. Nous pouvons nous demander comment ses génomes sont capables d'accumuler cet ADN tout en se préservant d'éventuels effets délétères ?

En étudiant 20 clades comprenant 168 espèces eucaryotes, Oliver et *al.* ont montré que le taux d'évolution de la taille du génome est proportionnelle à la taille du génome, avec les taux les plus rapides se produisant dans les plus grands génomes (Oliver et al., 2007). Ce modèle explique la répartition inégale de la taille des génomes chez les eucaryotes sans invoquer une force de sélection contre les grands génomes. Mais ce modèle n'est pas nécessairement applicable si la taille du génome atteint une valeur optimale spécifique à l'espèce déterminée par des variables écologiques et environnementales. Chez YAPH et YAYA, nous avons observé une tendance à la réduction du nombre de gènes en famille impliqués dans le métabolisme lipidique. Cette réduction du nombre de gènes est en corrélation avec les tailles de leur génome qui sont les plus petites du clade, avec 18,8 Mb pour YAYA et 16,2 Mb pour YAPH et aussi avec la taille moyenne de leurs régions intergèniques. Nous pourrions supposer que cette différence de taille ne provient pas d'une contraction après expansion mais simplement du fait que les génomes de YALI, YAGA, YAAL ont grandi de façon espèce spécifique. Ceci pourrait être en corrélation avec le plus fort pourcentage en GC de YALI, YAGA et YAAL qui serait potentiellement dû à la conversion génique, biaisée vers GC, sur les zones à réparer après les cassures menant aux réarrangements chromosomiques. Mais ceci ne corrèle pas avec les pertes de gènes et les pseudogénisations espèces spécifiques observées chez YAPH et YAYA grâce aux reconstitutions des scénarios évolutifs des familles impliquées dans le métabolisme lipidique. D'après la phylogénie des espèces et les reconstitutions de l'évolution des familles de gènes, l'expansion de la taille des génomes, comme l'expansion des familles de gènes, semble avoir démarré après la divergence de OHLI et avant celle de YAAL et semble se poursuivre chez YALI, YAGA et YAAL (4 duplications

YALI/YAGA spécifiques, 4 spécifiques à YALI, 1 à YAGA et 5 à YAAL). Ceci suggère que les génomes de YAPH et YAYA sont sous une pression de sélection les conduisant à éliminer les duplicats de gènes et à réduire les régions intergéniques. L'action de cette pression de sélection pourrait être corrélée à une nécessité à s'adapter à leur environnement pour par exemple diminuer le temps de génération ou la taille des cellules.

BIBLIOGRAPHIE

Acker, J., Ozanne, C., Kachouri-Lafond, R., Gaillardin, C., Neuvéglise, C., and Marck, C. (2008). Dicistronic tRNA-5S rRNA genes in Yarrowia lipolytica: an alternative TFIIIA-independent way for expression of 5S rRNA genes. Nucleic Acids Research 36, 5832-5844.

Ageitos, J., Vallejo, J., Veiga-Crespo, P., and Villa, T. (2011). Oily yeasts as oleaginous cell factories. Applied Microbiology and Biotechnology 90, 1219-1227.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of Molecular Biology 215, 403-410.

Alvarez, H., and Steinbüchel, A. (2002). Triacylglycerols in prokaryotic microorganisms. Applied Microbiology and Biotechnology 60, 367-376.

Anamnart, S., Tomita, T., Fukui, F., Fujimori, K., Harashima, S., Yamada, Y., and Oshima, Y. (1997). The P-OLE1 gene of Pichia angusta encodes a Δ-fatty acid desaturase and complements the ole1 mutation of Saccharomyces cerevisiae. Gene 184, 299-306.

Andrade, M.A., Perez-Iratxeta, C., and Ponting, C.P. (2001). Protein Repeats: Structures, Functions, and Evolution. Journal of Structural Biology 134, 117-131.

Athenstaedt, K. (2011). YALI0E32769g (DGA1) and YALI0E16797g (LRO1) encode major triacylglycerol synthases of the oleaginous yeast Yarrowia lipolytica. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids 1811, 587-596.

Athenstaedt, K., Jolivet, P., Boulard, C., Zivy, M., Negroni, L., Nicaud, J.-M., and Chardot, T. (2006). Lipid particle composition of the yeast Yarrowia lipolytica depends on the carbon source. PROTEOMICS 6, 1450-1459.

Athenstaedt, K., Zweytick, D., Jandrositz, A., Kohlwein, S.D., and Daum, G.n. (1999). Identification and Characterization of Major Lipid Particle Proteins of the Yeast Saccharomyces cerevisiae. Journal of Bacteriology 181, 6441-6448.

Babbitt, G.A., and Cotter, C.R. (2010). Functional Conservation of Nucleosome Formation Selectively Biases Presumably Neutral Molecular Variation in Yeast Genomes. Genome Biology and Evolution 3, 15-22.

Barth, G. (1985). Genetic regulation of isocitrate lyase in the yeast Yarrowia lipolytica. Current Genetics 10, 119-124.

Barth, G., and Gaillardin, C. (1996). Yarrowia lipolytica. In Non- conventional Yeasts in Biotechnology, K. Wolf, ed. (Berlin, Springer Verlag), pp. 313-388.

Barth, G., and Künkel, W. (1979). Alcohol dehydrogenase (ADH) in yeasts II. NAD+-and NADP+-dependent alcohol dehydrogenases in Saccharomycopsis lipolytica. Zeitschrift für allgemeine Mikrobiologie 19, 381-390.

Barth, G., Gaillardin, C., Mekouar, M., and Neuvéglise, C. (2013). Comparative Genomics of Yarrowia lipolytica. In Yarrowia lipolytica (Springer Berlin Heidelberg), pp. 1-30.

Bekpen, C., Marques-Bonet, T., Alkan, C., Antonacci, F., Leogrande, M.B., Ventura, M., Kidd, J.M., Siswara, P., Howard, J.C., and Eichler, E.E. (2009). Death and Resurrection of the Human IRGM Gene. PLoS Genet 5, e1000403.

Benjamin, S., and Pandey, A. (1998). Candida rugosa lipases: Molecular biology and versatility in biotechnology. Yeast 14, 1069-1087.

Beopoulos, A., Haddouche, R., Kabran, P., Dulermo, T., Chardot, T., and Nicaud, J.-M. (2012). Identification and characterization of DGA2, an acyltransferase of the DGAT1 acyl-CoA:diacylglycerol acyltransferase family in the oleaginous yeast Yarrowia lipolytica.

New insights into the storage lipid metabolism of oleaginous yeasts. Applied Microbiology and Biotechnology 93, 1523-1537.

Beopoulos, A., Mrozova, Z., Thevenieau, F., Le Dall, M.-T.r.s., Hapala, I., Papanikolaou, S., Chardot, T., and Nicaud, J.-M. (2008). Control of Lipid Accumulation in the Yeast Yarrowia lipolytica. Applied and Environmental Microbiology 74, 7779-7789.

Beopoulos, A., Nicaud, J.-M., and Gaillardin, C. (2011). An overview of lipid metabolism in yeasts and its impact on biotechnological processes. Applied Microbiology and Biotechnology 90, 1193-1206.

Beopoulos, A., Verbeke, J., Bordes, F., Guicherd, M., Bressy, M., Marty, A., and Nicaud, J.-M. (2014). Metabolic engineering for ricinoleic acid production in the oleaginous yeast Yarrowia lipolytica. Applied Microbiology and Biotechnology 98, 251-262.

Bergthorsson, U., Andersson, D.I., and Roth, J.R. (2007). Ohno's dilemma: Evolution of new genes under continuous selection. Proceedings of the National Academy of Sciences 104, 17004-17009.

Berninger, G., Schmidtchen, R., Casel, G., Knorr, A., Rautenstrauss, K., Kunau, W.-H., and Schweizer, E. (1993). Structure and metabolic control of the Yarrowia lipolytica peroxisomal 3-oxoacyl-CoA-thiolase gene. European Journal of Biochemistry 216, 607-613.

Bharti, R.K., Srivastava, S., and Thakur, I.S. (2014). Proteomic Analysis of Carbon Concentrating Chemolithotrophic Bacteria <italic>Serratia</italic> sp. for Sequestration of Carbon Dioxide. PLoS ONE 9, e91300.

Bierne, N., and Eyre-Walker, A. (2003). The Problem of Counting Sites in the Estimation of the Synonymous and Nonsynonymous Substitution Rates: Implications for the Correlation Between the Synonymous Substitution Rate and Codon Usage Bias. Genetics 165, 1587-1597.

Bigey, F., Tuery, K., Bougard, D., Nicaud, J.-M., and Moulin, G. (2003). Identification of a triacylglycerol lipase gene family in Candida deformans: molecular cloning and functional expression. Yeast 20, 233-248.

Bjorklund, A.K., Ekman, D., and Elofsson, A. (2006). Expansion of Protein Domain Repeats. PLoS Comput Biol 2, e114.

Bordes, F., Tarquis, L., Nicaud, J.-M., and Marty, A. (2011). Isolation of a thermostable variant of Lip2 lipase from Yarrowia lipolytica by directed evolution and deeper insight into the denaturation mechanisms involved. Journal of Biotechnology 156, 117-124.

Bosco, G., Campbell, P., Leiva-Neto, J.T., and Markow, T.A. (2007). Analysis of Drosophila Species Genome Size and Satellite DNA Content Reveals Significant Differences Among Strains as Well as Between Species. Genetics 177, 1277-1290.

Boulesteix, M., Weiss, M., and Biémont, C. (2006). Differences in Genome Size Between Closely Related Species: The Drosophila melanogaster Species Subgroup. Molecular Biology and Evolution 23, 162-167.

Boulton, C.A., and Ratledge, C. (1981). Correlation of Lipid Accumulation in Yeasts with Possession of ATP: Citrate Lyase. Journal of General Microbiology 127, 169-176.

Boutur, O., Dubreucq, E., and Galzy, P. (1995). Factors influencing ester synthesis catalysed in aqueous media by the lipase from candida deformans (zach) langeron and guerra. Journal of Biotechnology 42, 23-33.

Bregere, C., Rebrin, I., Sohal, R.S., Enrique, C., and Lester, P. (2008). Chapter 19 Detection and Characterization of In Vivo Nitration and Oxidation of Tryptophan Residues in Proteins. In Methods in Enzymology (Academic Press), pp. 339-349.

Bricker, D.K., Taylor, E.B., Schell, J.C., Orsak, T., Boutron, A., Chen, Y.-C., Cox, J.E., Cardon, C.M., Van Vranken, J.G., Dephoure, N., et al. (2012). A Mitochondrial Pyruvate Carrier Required for Pyruvate Uptake in Yeast, Drosophila, and Humans. Science 337, 96-100.

Brody, S., Oh, C., Hoja, U., and Schweizer, E. (1997). Mitochondrial acyl carrier protein is involved in lipoic acid synthesis in Saccharomyces cerevisiae. FEBS Letters 408, 217-220.

Butler, G., Rasmussen, M.D., Lin, M.F., Santos, M.A.S., Sakthikumar, S., Munro, C.A., Rheinbay, E., Grabherr, M., Forche, A., Reedy, J.L., et al. (2009). Evolution of pathogenicity and sexual reproduction in eight Candida genomes. Nature 459, 657-662.

Byrne, K.P., and Wolfe, K.H. (2005). The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Research 15, 1456-1461.

Casaregola, S., Neuvéglise, C., Lépingle, A., Bon, E., Feynerol, C., Artiguenave, F., Wincker, P., and Gaillardin, C. (2000). Genomic Exploration of the Hemiascomycetous Yeasts: 17. Yarrowia lipolytica. FEBS Letters 487, 95-100.

Castegna, A., Scarcia, P., Agrimi, G., Palmieri, L., Rottensteiner, H., Spera, I., Germinario, L., and Palmieri, F. (2010). Identification and Functional Characterization of a Novel Mitochondrial Carrier for Citrate and Oxoglutarate in Saccharomyces cerevisiae. Journal of Biological Chemistry 285, 17359-17370.

Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Molecular Biology and Evolution 17, 540-552.

Chang, C.-F., Chen, C.-C., Lee, C.-F., and Liu, S.-M. (2013). Identifying and characterizing Yarrowia keelungensis sp. nov., an oil-degrading yeast isolated from the sea surface microlayer. Antonie van Leeuwenhoek 104, 1117-1123.

Chen, B., Ling, H., and Chang, M.W. (2013). Transporter engineering for improved tolerance against alkane biofuels in *Saccharomyces cerevisiae*. Biotechnology for Biofuels 6, 21.

Chen, F.-C., and Li, W.-H. (2001). Genomic Divergences between Humans and Other Hominoids and the Effective Population Size of the Common Ancestor of Humans and Chimpanzees. American journal of human genetics 68, 444-456.

Choi, J.-Y., and Martin, C.E. (1999). The Saccharomyces cerevisiae FAT1 Gene Encodes an Acyl-CoA Synthetase That Is Required for Maintenance of Very Long Chain Fatty Acid Levels. Journal of Biological Chemistry 274, 4671-4683.

Claros, M.G., and Vincens, P. (1996). Computational Method to Predict Mitochondrially Imported Proteins and their Targeting Sequences. European Journal of Biochemistry 241, 779-786.

Colón, M., Hernàndez, F., López, K., Quezada, H., González, J., López, G., Aranda, C., and González, A. (2011). *Saccharomyces cerevisiae* Bat1 and Bat2 Aminotransferases Have Functionally Diverged from the Ancestral-Like Kluyveromyces lactis Orthologous Enzyme. PLoS ONE 6, e16099.

Contreras-Shannon, V., Lin, A.-P., McCammon, M.T., and McAlister-Henn, L. (2005). Kinetic Properties and Metabolic Contributions of Yeast Mitochondrial and Cytosolic NADP+ specific Isocitrate Dehydrogenases. Journal of Biological Chemistry 280, 4469-4475.

Czabany, T., Athenstaedt, K., and Daum, G. (2007). Synthesis, storage and degradation of neutral lipids in yeast. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids 1771, 299-309.

Darvishi, F. (2012). Expression of native and mutant extracellular lipases from *Yarrowia lipolytica* in *Saccharomyces cerevisiae*. Microbial Biotechnology 5, 634-641.

Darvishi, F., Destain, J., Nahvi, I., Thonart, P., and Zarkesh-Esfahani, H. (2011). High-level production of extracellular lipase by *Yarrowia lipolytica* mutants from methyl oleate. New Biotechnology 28, 756-760.

Das, S., Roymondal, U., and Sahoo, S. (2009). Analyzing gene expression from relative codon usage bias in Yeast genome: A statistical significance and biological relevance. Gene 443, 121-131.

de Almeida, A.F., Tauk-Tornisielo, S.M., and Carmona, E.C. (2013). Acid Lipase from *Candida viswanathii*: Production, Biochemical Properties, and Potential Application. BioMed Research International 2013, 10.

De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006). CAFE: a computational tool for the study of gene family evolution. Bioinformatics 22, 1269-1271.

De Hertogh, B.Æ., Hancy, F.d.r., Goffeau, A., and Baret, P.V. (2006). Emergence of Species-Specific Transporters During Evolution of the *Hemiascomycete Phylum*. Genetics 172, 771-781.

Des Marais, D.L., and Rausher, M.D. (2008). Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature 454, 762-765.

Dmochowska, A., Dignard, D., Maleszka, R., and Thomas, D.Y. (1990). Structure and transcriptional control of the *Saccharomyces cerevisiae* POX1 gene encoding acylcoenzyme A oxidase. Gene 88, 247-252.

Doroftei, A. (2012). Évolution de familles de gènes par duplications et pertes : algorithmes pour la correction d'arbres bruités. In Informatique (Montréal, Université de Montréal).

Drillon, G., Carbone, A., and Fischer, G. (2014). SynChro: A Fast and Easy Tool to Reconstruct and Visualize Synteny Blocks along Eukaryotic Chromosomes. PLoS ONE 9, e92621.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., de Montigny, J., Marck, C., Neuveglise, C., Talla, E., et al. (2004). Genome evolution in yeasts. Nature 430, 35 - 44.

Dulermo, T., and Nicaud, J.-M. (2011). Involvement of the G3P shuttle and ß-oxidation pathway in the control of TAG synthesis and lipid accumulation in *Yarrowia lipolytica*. Metabolic Engineering 13, 482-491.

Dulermo, T., Tréton, B., Beopoulos, A., Kabran Gnankon, A.P., Haddouche, R., and Nicaud, J.-M. (2013). Characterization of the two intracellular lipases of *Y. lipolytica* encoded by TGL3 and TGL4 genes: New insights into the role of intracellular lipases and lipid body organisation. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids 1831, 1486-1495.

Einerhand, A.W.C., Voorn-Brouwer, T.M., Erdmann, R., Kunau, W.-H., and Tabak, H.F. (1991). Regulation of transcription of the gene coding for peroxisomal 3-oxoacyl-CoA thiolase of *Saccharomcyes cerevisiae*. European Journal of Biochemistry 200, 113-122.

Eirich, L.D., Craft, D.L., Steinberg, L., Asif, A., Eschenfeldt, W.H., Stols, L., Donnelly, M.I., and Wilson, C.R. (2004). Cloning and Characterization of Three Fatty Alcohol Oxidase Genes from *Candida tropicalis* Strain ATCC 20336. Applied and Environmental Microbiology 70, 4872-4879.

Farooq, M.A., Pracheil, T.M., Dong, Z., Xiao, F., and Liu, Z. (2013). Mitochondrial DNA Instability in Cells Lacking Aconitase Correlates with Iron Citrate Toxicity. Oxidative Medicine and Cellular Longevity 2013, 10.

Fedorova, N.D., Khaldi, N., Joardar, V.S., Maiti, R., Amedeo, P., Anderson, M.J., Crabtree, J., Silva, J.C., Badger, J.H., Albarraq, A., et al. (2008). Genomic Islands in the Pathogenic Filamentous Fungus *Aspergillus fumigatus*. PLoS Genet 4, e1000046.

Fickers, P., Benetti, P.H., Waché, Y., Marty, A., Mauersberger, S., Smit, M.S., and Nicaud, J.M. (2005a). Hydrophobic substrate utilisation by the yeast *Yarrowia lipolytica*, and its potential applications. FEMS Yeast Research 5, 527-543.

Fickers, P., Fudalej, F., Dall, M.T.L., Casaregola, S., Gaillardin, C., Thonart, P., and Nicaud, J.M. (2005b). Identification and characterisation of LIP7 and LIP8 genes encoding two extracellular triacylglycerol lipases in the yeast *Yarrowia lipolytica*. Fungal Genetics and Biology 42, 264-274.

Fickers, P., Marty, A., and Nicaud, J.M. (2011). The lipases from *Yarrowia lipolytica*: Genetics, production, regulation, biochemical characterization and biotechnological applications. Biotechnology Advances 29, 632-644.

Fink, G. (1987). Pseudogenes in yeast? Cell 49, 5 - 6.

Fischer, G., James, S.A., Roberts, I.N., Oliver, S.G., and Louis, E.J. (2000). Chromosomal evolution in Saccharomyces. Nature 405, 451-454.

Fischer, G., Neuvéglise, C., Durrens, P., Gaillardin, C., and Dujon, B. (2001). Evolution of Gene Order in the Genomes of Two Related Yeast Species. Genome Research 11, 2009-2019.

Fitch, W.M. (1971). Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. Systematic Zoology 20, 406-416.

Fitch, W.M. (2000). Homology: a personal view on some of the problems. Trends in genetics : TIG 16, 227-231.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.-l., and Postlethwait, J. (1999). Preservation of Duplicate Genes by Complementary, Degenerative Mutations. Genetics 151, 1531-1545.

Franken, J., Kroppenstedt, S., Swiegers, J., and Bauer, F. (2008). Carnitine and carnitine acetyltransferases in the yeast *Saccharomyces cerevisiae*: a role for carnitine in stress protection. Current Genetics 53, 347-360.

Frenkel, Z.M., Barzily, Z., Volkovich, Z., and Trifonov, E.N. (2013). Hidden ancient repeats in DNA: Mapping and quantification. Gene 528, 282-287.

Friedrich, A., Jung, P.P., Hou, J., Neuvéglise, C., and Schacherer, J. (2012). Comparative Mitochondrial Genomics within and among Yeast Species of the *Lachancea* Genus. PLoS ONE *7*, e47834.

Gabaldon, T., Martin, T., Marcet-Houben, M., Durrens, P., Bolotin-Fukuhara, M., Lespinet, O., Arnaise, S., Boisnard, S., Aguileta, G., Atanasova, R., et al. (2013). Comparative genomics of emerging pathogens in the *Candida glabrata* clade. BMC Genomics 14, 623.

Gao, L.-Z., and Innan, H. (2004). Very low gene duplication rate in the yeast genome. Science 306, 1367 - 1370.

Gardini, F., Tofalo, R., Belletti, N., Iucci, L., Suzzi, G., Torriani, S., Guerzoni, M.E., and Lanciotti, R. (2006). Characterization of yeasts involved in the ripening of *Pecorino Crotonese* cheese. Food Microbiology 23, 641-648.

Gavin, C.C., and Kenneth, H.W. (2007). Increased glycolytic flux as an outcome of whole-genome duplication in yeast. Molecular Systems Biology 3.

Glover, J.R., Andrews, D.W., and Rachubinski, R.A. (1994). *Saccharomyces cerevisiae* peroxisomal thiolase is imported as a dimer. Proceedings of the National Academy of Sciences 91, 10541-10545.

Goffeau, A., Barrell, B., Bussey, H., Davis, R., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J., Jacq, C., Johnston, M., et al. (1996). Life with 6000 genes. Science 274, 546 - 567.

Gojobori, T., Li, W.-H., and Graur, D. (1982). Patterns of nucleotide substitution in pseudogenes and functional genes. Journal of Molecular Evolution 18, 360-369.

Goldman, N., and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution 11, 725-736.

Goldsmith, M.E., Humphries, R.K., Ley, T., Cline, A., Kantor, J.A., and Nienhuis, A.W. (1983). "Silent" nucleotide substitution in a beta+-thalassemia globin gene activates splice site in coding sequence RNA. Proceedings of the National Academy of Sciences 80, 2318-2322.

Goodman, M., Czelusniak, J., Moore, G.W., Romero-Herrera, A.E., and Matsuda, G. (1979). Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences. Systematic Biology 28, 132-163.

Graybill, E.R., Rouhier, M.F., Kirby, C.E., and Hawes, J.W. (2007). Functional comparison of citrate synthase isoforms from *S. cerevisiae*. Archives of Biochemistry and Biophysics 465, 26-37.

Gregory, T.R. (2001). The Bigger the C-Value, the Larger the Cell: Genome Size and Red Blood Cell Size in Vertebrates. Blood Cells, Molecules, and Diseases 27, 830-843.

Groenewald, M., and Smith, M. (2013). The teleomorph state of *Candida deformans* Langeron & Guerra and description of *Yarrowia yakushimensis* comb. nov. Antonie van Leeuwenhoek 103, 1023-1028.

Groenewald, M., Boekhout, T., Neuvéglise, C., Gaillardin, C., van Dijck, P.W.M., and Wyss, M. (2014). *Yarrowia lipolytica*: Safety assessment of an oleaginous yeast with a great industrial potential. Critical Reviews in Microbiology 40, 187-206.

Guillaume, B. (2007). Comparing Genomes with Duplications: A Computational Complexity Point of View. IEEE/ACM Transactions on Computational Biology and Bioinformatics 4, 523-534.

Guindon, S.p., Dufayard, J.-F.o., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology 59, 307-321.

Guo, T., Kit, Y.Y., Nicaud, J.-M., Le Dall, M.-T., Sears, S.K., Vali, H., Chan, H., Rachubinski, R.A., and Titorenko, V.I. (2003). Peroxisome division in the yeast *Yarrowia lipolytica* is regulated by a signal from inside the peroxisome. The Journal of Cell Biology 162, 1255-1266.

Haddouche, R., Delessert, S., Sabirova, J., Neuvéglise, C., Poirier, Y., and Nicaud, J.-M. (2010). Roles of multiple acyl-CoA oxidases in the routing of carbon flow towards β-oxidation and polyhydroxyalkanoate biosynthesis in *Yarrowia lipolytica*. FEMS Yeast Research 10, 917-927.

Haddouche, R., Poirier, Y., Delessert, S., Sabirova, J., Pagot, Y., Neuvéglise, C., and Nicaud, J.-M. (2011). Engineering polyhydroxyalkanoate content and monomer composition in the oleaginous yeast *Yarrowia lipolytica* by modifying the ß-oxidation multifunctional protein. Applied Microbiology and Biotechnology 91, 1327-1340.

Hahn, M., Han, M., and Han, S. (2007). Gene family evolution across 12 Drosophila genomes. PLoS Genet 3, e197.

Hane, J.K., Rouxel, T., Howlett, B.J., Kema, G.H., Goodwin, S.B., and Oliver, R.P. (2011). A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. Genome Biol 12, R45.

Harris, R., Paxton, R., Powell, S., Goodwin, G., Kuntz, M., and Han, A. (1986). Regulation of branched-chain alpha-ketoacid dehydrogenase complex by covalent modification. Adv Enzyme Regul 25, 219-237.

Harris, R.A., Joshi, M., Jeoung, N.H., and Obayashi, M. (2005). Overview of the Molecular and Biochemical Basis of Branched-Chain Amino Acid Catabolism. The Journal of Nutrition 135, 1527S-1530S.

Harrison, P., Kumar, A., Lan, N., Echols, N., Snyder, M., and Gerstein, M. (2002). A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. Journal of Molecular Biology 316, 409-419.

Haselbeck, R.J., and McAlister-Henn, L. (1991). Isolation, nucleotide sequence, and disruption of the Saccharomyces cerevisiae gene encoding mitochondrial NADP(H)-specific isocitrate dehydrogenase. Journal of Biological Chemistry 266, 2339-2345.

Henke, B., Girzalsky, W., Berteaux-Lecellier, V., and Erdmann, R. (1998). IDP3 Encodes a Peroxisomal NADP-dependent Isocitrate Dehydrogenase Required for the ß-Oxidation of Unsaturated Fatty Acids. Journal of Biological Chemistry 273, 3702-3711.

Herzig, S., Raemy, E., Montessuit, S., Veuthey, J.-L., Zamboni, N., Westermann, B., Kunji, E.R.S., and Martinou, J.-C. (2012). Identification and Functional Expression of the Mitochondrial Pyruvate Carrier. Sciences 337, 93-96.

Hettema, E.H., and Tabak, H.F. (2000). Transport of fatty acids and metabolites across the peroxisomal membrane. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids 1486, 18-27.

Hiltunen, J.K., Mursula, A.M., Rottensteiner, H., Wierenga, R.K., Kastaniotis, A.J., and Gurvitz, A. (2003). The biochemistry of peroxisomal ß-oxidation in the yeast *Saccharomyces cerevisiae*. FEMS Microbiology Reviews 27, 35-64.

Hiltunen, J.K., Wenzel, B., Beyer, A., Erdmann, R., Fossa, A., and Kunau, W.H. (1992). Peroxisomal multifunctional beta-oxidation protein of *Saccharomyces cerevisiae*. Molecular analysis of the fox2 gene and gene product. Journal of Biological Chemistry 267, 6646-6653.

Hirakawa, K., Kobayashi, S., Inoue, T., Endoh-Yamagami, S., Fukuda, R., and Ohta, A. (2009). Yas3p, an Opi1 Family Transcription Factor, Regulates Cytochrome P450 Expression in Response to n-Alkanes in *Yarrowia lipolytica*. Journal of Biological Chemistry 284, 7126-7137.

Hiraoka, Y., Kawamata, K., Haraguchi, T., and Chikashige, Y. (2009). Codon usage bias is correlated with gene expression levels in the fission yeast *Schizosaccharomyces pombe*. Genes to Cells 14, 499-509.

Holz, M., Förster, A., Mauersberger, S., and Barth, G. (2009). Aconitase overexpression changes the product ratio of citric acid production by *Yarrowia lipolytica*. Applied Microbiology and Biotechnology 81, 1087-1096.

Holz, M., Otto, C., Kretzschmar, A., Yovkova, V., Aurich, A., Petter, M., Marx, A., and Barth, G. (2011). Overexpression of alpha-ketoglutarate dehydrogenase in *Yarrowia lipolytica* and its effect on production of organic acids. Applied Microbiology and Biotechnology 89, 1519-1526.

Huang, C., Chen, X.-F., Yang, X.-Y., Xiong, L., Lin, X.-Q., Yang, J., Wang, B., and Chen, X.-D. (2014). Bioconversion of Corncob Acid Hydrolysate into Microbial Oil by the Oleaginous Yeast *Lipomyces starkeyi*. Applied Biochemistry and Biotechnology 172, 2197-2204.

Hube, B., Stehr, F., Bossenz, M., Mazur, A., Kretschmar, M., and Schäfer, W. (2000). Secreted lipases of *Candida albicans*: cloning, characterisation and expression analysis of a new gene family with at least ten members. Archives of Microbiology 174, 362-374.

Hughes, A.L. (1994). The Evolution of Functionally Novel Proteins after Gene Duplication. Proceedings of the Royal Society of London Series B: Biological Sciences 256, 119-124.

Hughes, A.L., and Friedman, R. (2003). Parallel Evolution by Gene Duplication in the Genomes of Two Unicellular Fungi. Genome Research 13, 794-799.

Hughes, T.R., Roberts, C.J., Dai, H., Jones, A.R., Meyer, M.R., Slade, D., Burchard, J., Dow, S., Ward, T.R., Kidd, M.J., et al. (2000). Widespread aneuploidy revealed by DNA microarray expression profiling. Nat Genet 25, 333-337.

Hynes, M.J., and Murray, S.L. (2010). ATP-Citrate Lyase Is Required for Production of Cytosolic Acetyl Coenzyme A and Development in *Aspergillus nidulans*. Eukaryotic Cell 9, 1039-1048.

Iida, T., Ohta, A., and Takagi, M. (1998). Cloning and characterization of an n-alkane-inducible cytochrome P450 gene essential for n-decane assimilation by *Yarrowia lipolytica*. Yeast 14, 1387-1397.

Iida, T., Sumita, T., Ohta, A., and Takagi, M. (2000). The cytochrome P450ALK multigene family of an n-alkane-assimilating yeast, *Yarrowia lipolytica*: cloning and characterization of genes coding for new CYP52 family members. Yeast 16, 1077-1087.

Ina, Y. (1995). New methods for estimating the numbers of synonymous and nonsynonymous substitutions. Journal of Molecular Evolution 40, 190-226.

Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11, 97-108.

Ismail, S.A.S., Deak, T., Abd El-Rahman, H.A., Yassien, M.A.M., and Beuchat, L.R. (2001). Effectiveness of immersion treatments with acids, trisodium phosphate, and herb decoctions in reducing populations of *Yarrowia lipolytica* and naturally occurring aerobic microorganisms on raw chicken. International Journal of Food Microbiology 64, 13-19.

Jackson, A.P., Gamble, J.A., Yeomans, T., Moran, G.P., Saunders, D., Harris, D., Aslett, M., Barrell, J.F., Butler, G., Citiulo, F., *et al.* (2009). Comparative genomics of the fungal pathogens *Candida dubliniensis* and *Candida albicans*. Genome Research *19*, 2231-2244.

Jain, M.R., Zinjarde, S.S., Deobagkar, D.D., and Deobagkar, D.N. (2004). 2,4,6-Trinitrotoluene transformation by a tropical marine yeast, *Yarrowia lipolytica* NCIM 3589. Marine Pollution Bulletin 49, 783-788.

Jensen, R. (2001). Orthologs and paralogs - we need to get it right. Genome Biology 2, interactions1002.1001 - interactions1002.1003.

Jia, Y.K., Bécam, A.M., and Herbert, C.J. (1997). The CIT3 gene of *Saccharomyces cerevisiae* encodes a second mitochondrial isoform of citrate synthase. Molecular Microbiology 24, 53-59.

Kabran, P. (2011). Lipid metabolism and glyoxylate cycle in the yeast *Yarrowia lipolytica* (Agriculture, alimentation, biologie, environnement et santé (ABIES) - ED 435).

Kabran, P., Rossignol, T., Gaillardin, C., Nicaud, J.-M., and Neuvéglise, C. (2012). Alternative Splicing Regulates Targeting of Malate Dehydrogenase in *Yarrowia lipolytica*. DNA Research 19, 231-244.

Kaessmann, H., Vinckenbosch, N., and Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet 10, 19-31.

Katju, V., Farslow, J., and Bergthorsson, U. (2009). Variation in gene duplicates with low synonymous divergence in *Saccharomyces cerevisiae* relative to *Caenorhabditis elegans*. Genome Biology 10, R75.

Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research *30*, 3059-3066.

Kawasaki, Y., and Sugino, A. (2001). Yeast replicative DNA polymerases and their role at the replication fork. Anglais 12, 9.

Keeling, P.J., Burger, G., Durnford, D.G., Lang, B.F., Lee, R.W., Pearlman, R.E., Roger, A.J., and Gray, M.W. (2005). The tree of eukaryotes. Trends in ecology & evolution (Personal edition) 20, 670-676.

Keller, I., Bensasson, D., and Nichols, R.A. (2007). Transition-Transversion Bias Is Not Universal: A Counter Example from Grasshopper Pseudogenes. PLoS Genet 3, e22.

Kellis, M., Birren, B.W., and Lander, E.S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 428, 617-624.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423, 241 - 254.

Kimura, M., and Ohta, T. (1974). On Some Principles Governing Molecular Evolution. Proceedings of the National Academy of Sciences 71, 2848-2852.

Knutsen, A.K., Robert, V., Poot, G.A., Epping, W., Figge, M., Holst-Jensen, A., Skaar, I., and Smith, M.T. (2007). Polyphasic re-examination of *Yarrowia lipolytica* strains and the description of three novel *Candida* species: *Candida oslonensis* sp. nov., *Candida alimentaria* sp. nov. and *Candida hollandica* sp. nov. Int J Syst Evol Microbiol 57, 2426-2435.

Koei, O., Hironobu, T., Sakuzo, F., Ichiro, K., and Tatsuyuki, K. (1987). Peroxisomal acyl-coenzyme A oxidase multigene family of the yeast *Candida tropicalis*; nucleotide sequence of a third gene and its protein product. Gene 58, 37-44.

Koivistoinen, O., Kuivanen, J., Barth, D., Turkia, H., Pitkanen, J.-P., Penttila, M., and Richard, P. (2013). Glycolic acid production in the engineered yeasts *Saccharomyces cerevisiae* and *Kluyveromyces lactis*. Microbial Cell Factories 12, 82.

Kosíková, B., and Sláviková, E. (1996). Growth of *Saccharomyces cerevisiae*, *Rhodotorula rubra* and *Bullera alba* in the presence of beechwood prehydrolyzate-based lignin fractions. Folia microbiologica 41, 430-432.

Koszul, R., and Fischer, G. (2009). A prominent role for segmental duplications in modeling Eukaryotic genomes. Comptes Rendus Biologies 332, 254-266.

Kozłowski, J., Konarzewski, M., and Gawelczyk, A.T. (2003). Cell size as a link between noncoding DNA and metabolic rate scaling. Proceedings of the National Academy of Sciences 100, 14080-14085.

Kubicek, C., Baker, S., Gamauf, C., Kenerley, C., and Druzhinina, I. (2008). Purifying selection and birth-and-death evolution in the class II hydrophobin gene families of the ascomycete Trichoderma/Hypocrea. BMC Evolutionary Biology 8, 4.

Kumari, A., and Gupta, R. (2012). Extracellular expression and characterization of thermostable lipases, LIP8, LIP14 and LIP18, from *Yarrowia lipolytica*. Biotechnology Letters 34, 1733-1739.

Kumari, A., and Gupta, R. (2013). Phenylalanine to leucine point mutation in oxyanion hole improved catalytic efficiency of Lip12 from *Yarrowia lipolytica*. Enzyme and Microbial Technology 53, 386-390.

Kumari, A., Verma, V., and Gupta, R. (2012). Comparative biochemical characterization and in silico analysis of novel lipases Lip11 and Lip12 with Lip2 from *Yarrowia lipolytica*. World Journal of Microbiology and Biotechnology 28, 3103-3111.

Kursu, V.A.S., Pietikäinen, L.P., Fontanesi, F., Aaltonen, M.J., Suomi, F., Raghavan Nair, R., Schonauer, M.S., Dieckmann, C.L., Barrientos, A., Hiltunen, J.K., et al. (2013). Defects in mitochondrial fatty acid synthesis result in failure of multiple aspects of mitochondrial biogenesis in *Saccharomyces cerevisiae*. Molecular Microbiology 90, 824-840.

Kurtz, S., Phillippy, A., Delcher, A., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. (2004). Versatile and open software for comparing large genomes. Genome Biology *5*, R12.

Kurtzman, C.P. (2005). New species and a new combination in the *Hyphopichia* and *Yarrowia* yeast clades. Antonie van Leeuwenhoek 88, 121-130.

Kurtzman, C.P. (2011a). Discussion of Teleomorphic and Anamorphic Ascomycetous Yeasts and Yeast-like Taxa. In The Yeasts: a taxonomic study C.P. Kurtzman, J.W. Fell, and T. Boekhout, eds. (Elsevier), pp. 293-307.

Kurtzman, C.P. (2011b). Yarrowia van der Walt & von Arx (1980). In The Yeasts: a taxonomic study C.P. Kurtzman, J.W. Fell, and T. Boekhout, eds. (Elsevier), pp. 927-930.

Kurtzman, C.P., and Fell, J.W. (1998). The Yeasts: a taxonomic study 4edn (Amsterdam, Elsevier).

Lachance, M.A., Boekhout, T., Scorzetti, J.W.F., and Kurtzman, C.P. (2011). Candida Berkhout (1923). In The Yeasts: a taxonomic study C.P. Kurtzman, J.W. Fell, and T. Boekhout, eds. (Elsevier), pp. 987-1278.

Lafontaine, I., and Dujon, B. (2010). Origin and fate of pseudogenes in Hemiascomycetes: a comparative analysis. BMC Genomics 11, 260.

Larroy, C., Rosario Fernandez, M., Gonzalez, E., Parés, X., and Biosca, J.A. (2003). Properties and functional significance of *Saccharomyces cerevisiae* ADHVI. Chemico-Biological Interactions 143,Äì144, 229-238.

Lawson, J.E., Behal, R.H., and Reed, L.J. (1991). Disruption and mutagenesis of the *Saccharomyces cerevisiae* PDX1 gene encoding the protein X component of the pyruvate dehydrogenase complex. Biochemistry 30, 2834-2839.

Lazarow, P.B. (2006). Chapter 3.1.7. The import receptor Pex7p and the PTS2 targeting sequence. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research 1763, 1599-1604.

Leber, R., Zinser, E., Zellnig, G., Paltauf, F., and Daum, G. (1994). Characterization of lipid particles of the yeast, *Saccharomyces cerevisiae*. Yeast 10, 1421-1428.

Lennzrz, W.J. (1961). The role of isoleucine in the biosynthesis of branched-chain fatty acids by micrococcus lysodeikticus. Biochemical and Biophysical Research Communications 6, 112-116.

Lewin, A.S., Hines, V., and Small, G.M. (1990). Citrate synthase encoded by the CIT2 gene of *Saccharomyces cerevisiae* is peroxisomal. Molecular and Cellular Biology 10, 1399-1405.

Limtong, S., Youngmanitchai, W., Kawasaki, H., and Seki, T. (2008). *Candida phangngensis* sp. nov., an anamorphic yeast species in the *Yarrowia* clade, isolated from water in mangrove forests in Phang-Nga Province, Thailand. Int J Syst Evol Microbiol 58, 515-519.

Ling, H., Chen, B., Kang, A., Lee, J.-M., and Chang, M. (2013). Transcriptome response to alkane biofuels in *Saccharomyces cerevisiae*: identification of efflux pumps involved in alkane tolerance. Biotechnology for Biofuels 6, 95.

Liti, G., Carter, D.M., Moses, A.M., Warringer, J., Parts, L., James, S.A., Davey, R.P., Roberts, I.N., Burt, A., Koufopanou, V., et al. (2009). Population genomics of domestic and wild yeasts. Nature 458, 337-341.

Liu, W.-S., Pan, X.-X., Jia, B., Zhao, H.-Y., Xu, L., Liu, Y., and Yan, Y.-J. (2010). Surface display of active lipases Lip7 and Lip8 from *Yarrowia Lipolytica* on *Saccharomyces Cerevisiae*. Applied Microbiology and Biotechnology 88, 885-891.

Luo, Y.-S., Nicaud, J.-M., Van Veldhoven, P.P., and Chardot, T. (2002). The acyl-CoA oxidases from the yeast *Yarrowia lipolytica*: characterization of Aox2p. Archives of Biochemistry and Biophysics 407, 32-38.

Luo, Y.-S., Wang, H.-J., Gopalan, K.V., Srivastava, D.K., Nicaud, J.-M., and Chardot, T. (2000). Purification and Characterization of the Recombinant Form of Acyl CoA Oxidase 3 from the Yeast *Yarrowia lipolytica*. Archives of Biochemistry and Biophysics 384, 1-8.

Lynch, M., and Conery, J. (2000). The evolutionary fate and consequences of duplicate genes. Science 290, 1151 - 1155.

Lynch, M., and Conery, J. (2003a). The evolutionary demography of duplicate genes. In Genome Evolution (Springer Netherlands), pp. 35-44.

Lynch, M., and Conery, J. (2003b). The origins of genome complexity. Science 302, 1401 - 1404.

Maclean, C.J., and Greig, D. (2011). Reciprocal gene loss following experimental whole-genome duplication causes reproductive isolation in yeast. Evolution 65, 932-945.

Maeda, I., Delessert, S., Hasegawa, S., Seto, Y., Zuber, S., and Poirier, Y. (2006). The Peroxisomal Acyl-CoA Thioesterase Pte1p from *Saccharomyces cerevisiae* Is Required for Efficient Degradation of Short Straight Chain and Branched Chain Fatty Acids. Journal of Biological Chemistry 281, 11729-11735.

Maeda, N., and Smithies, O. (1986). The Evolution of Multigene Families: Human Haptoglobin Genes. Annual Review of Genetics 20, 81-108.

Marcotte, E.M., Pellegrini, M., Yeates, T.O., and Eisenberg, D. (1999). A census of protein repeats. Journal of Molecular Biology 293, 151-160.

Marques, A., Vinckenbosch, N., Brawand, D., and Kaessmann, H. (2008). Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. Genome Biology 9, R54.

Masuda, Y., Park, S.M., Ohta, A., and Takagi, M. (1995). Cloning and characterization of the POX2 gene in *Candida maltosa*. Gene 167, 157-161.

Matatiele, P.R. (2005). Fatty alcohol and fatty aldehyde dehydrogenases of *Yarrowia lipolytica* (Bloemfontein, South Africa, University of the Free State).

Maxwell, P.H., and Curcio, M.J. (2007). Retrosequence formation restructures the yeast genome. Genes & Development 21, 3308-3318.

McAlister-Henn, L., and Thompson, L.M. (1987). Isolation and expression of the gene encoding yeast mitochondrial malate dehydrogenase. Journal of Bacteriology 169, 5157-5166.

McDonald, J.H., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature 351, 652-654.

Mekouar, M., Blanc-Lenfle, I., Ozanne, C., Da Silva, C., Cruaud, C., Wincker, P., Gaillardin, C., and Neuveglise, C. (2010). Detection and analysis of alternative splicing in *Yarrowia lipolytica* reveal structural constraints facilitating nonsense-mediated decay of intron-retaining transcripts. Genome Biology 11, R65.

Merico, A., Sulo, P., Piškur, J., and Compagno, C. (2007). Fermentative lifestyle in yeasts belonging to the Saccharomyces complex. FEBS Journal 274, 976-989.

Michely, S., Devillers, H., Sarilar, V., Dulermo, R., Crutz-Le Coq, A.-M., Leh-Louis, V., Nicolas, M., Nguyen, H.-V., Juliette, P., Jossinet, F., et al. (in prep.). Whole genome sequencing as an help to reclassify *Candida* species: the case of *Oleozyma (Candida) hispaniensis*.

Michely, S., Toulza, E., Subirana, L., John, U., Cognat, V., Maréchal-Drouard, L., Grimsley, N., Moreau, H., and Piganeau, G. (2013). Evolution of Codon Usage in the Smallest

Photosynthetic Eukaryotes and Their Giant Viruses. Genome Biology and Evolution 5, 848-859.

Minard, K.I., and McAlister-Henn, L. (1991). Isolation, nucleotide sequence analysis, and disruption of the MDH2 gene from *Saccharomyces cerevisiae*: evidence for three isozymes of yeast malate dehydrogenase. Molecular and Cellular Biology 11, 370-380.

Miran, S.G., Lawson, J.E., and Reed, L.J. (1993). Characterization of PDH beta 1, the structural gene for the pyruvate dehydrogenase beta subunit from *Saccharomyces cerevisiae*. Proceedings of the National Academy of Sciences 90, 1252-1256.

Miyabe, I., Kunkel, T.A., and Carr, A.M. (2011). The Major Roles of DNA Polymerases Epsilon and Delta at the Eukaryotic Replication Fork Are Evolutionarily Conserved. PLoS Genet 7, e1002407.

Mizugaki, M., Koeda, N., Kondo, A., Kimura, C., and Yamanaka, H. (1985). Studies on the Metabolism of Unsaturated Fatty Acids. XVI. Purification and General Properties of 2, 4-Dienoyl-CoA Reductase from *Candida lipolytica*. Journal of Biochemistry 97, 837-843.

Mlíčková, K., Roux, E., Athenstaedt, K., d'Andrea, S., Daum, G., Chardot, T., and Nicaud, J.-M. (2004). Lipid Accumulation, Lipid Body Formation, and Acyl Coenzyme A Oxidases of the Yeast *Yarrowia lipolytica*. Applied and Environmental Microbiology 70, 3918-3924.

Morales, L., and Dujon, B. (2012). Evolutionary Role of Interspecies Hybridization and Genetic Exchanges in Yeasts. Microbiology and Molecular Biology Reviews 76, 721-739.

Morin, N., Cescut, J., Beopoulos, A., Lelandais, G., Le Berre, V., Uribelarrea, J.-L., Molina-Jouve, C., and Nicaud, J.-M. (2011). Transcriptomic Analyses during the Transition from Biomass Production to Lipid Accumulation in the Oleaginous Yeast *Yarrowia lipolytica*. PLoS ONE 6, e27966.

Morris, R.T., and Drouin, G. (2011). Ectopic Gene Conversions in the Genome of Ten Hemiascomycete Yeast Species. International Journal of Evolutionary Biology 2011, 11.

Muderhwa, J.M., Ratomahenina, R., Pina, M., Graille, J., and Galzy, P. (1985). Purification and properties of the lipase from *Candida deformans* (Zach) Langeron and Guerra, Vol 62 (Heidelberg, ALLEMAGNE, Springer).

Nagy, E., Dlauchy, D., Medeiros, A., Péter, G., and Rosa, C. (2014). *Yarrowia porcina* sp. nov. and *Yarrowia bubula* f.a. sp. nov., two yeast species from meat and river sediment. Antonie van Leeuwenhoek 105, 697-707.

Nagy, E., Niss, M., Dlauchy, D., Arneborg, N., Nielsen, D.S., and Péter, G. (2013). *Yarrowia divulgata* f.a., sp. nov., a yeast species from animal-related and marine sources. International Journal of Systematic and Evolutionary Microbiology 63, 4818-4823.

Nagylaki, T. (1984). Evolution of multigene families under interchromosomal gene conversion. Proceedings of the National Academy of Sciences 81, 3796-3800.

Näsvall, J., Sun, L., Roth, J.R., and Andersson, D.I. (2012). Real-Time Evolution of New Genes by Innovation, Amplification, and Divergence. Science 338, 384-387.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3, 418 - 426.

Nei, M., and Rooney, A.P. (2005). Concerted and birth-and-death evolution of multigene families. Annual Review of Genetics 39, 121-152.

Nei, M., Rogozin, I.B., and Piontkivska, H. (2000). Purifying selection and birth-and-death evolution in the ubiquitin gene family. Proceedings of the National Academy of Sciences 97, 10866-10871.

Nikolski, M., and Sherman, D.J. (2007). Family relationships: should consensus reign? - consensus clustering for protein families. Bioinformatics 23, e71-e76.

Niu, X.D., Browning, K.S., Behal, R.H., and Reed, L.J. (1988). Cloning and nucleotide sequence of the gene for dihydrolipoamide acetyltransferase from *Saccharomyces cerevisiae*. Proceedings of the National Academy of Sciences 85, 7546-7550.

Nyns, E.J., Chiang, N., and Wiaux, A.L. (1968). Comparative lipid content of *Candida lipolytica* grown on glucose and onn-hexadecane. Antonie van Leeuwenhoek 34, 197-204.

Obermeyer, T., Fraisl, P., DiRusso, C.C., and Black, P.N. (2007). Topology of the yeast fatty acid transport protein Fat1p: mechanistic implications for functional domains on the cytosolic surface of the plasma membrane. Journal of Lipid Research 48, 2354-2364.

Ohno, S. (1970). Evolution by gene duplication (Springer-Verlag).

Oliver, M.J., Petrov, D., Ackerly, D., Falkowski, P., and Schofield, O.M. (2007). The mode and tempo of genome size evolution in eukaryotes. Genome Research 17, 594-601.

Olson, M.V. (1999). When Less Is More: Gene Loss as an Engine of Evolutionary Change. The American Journal of Human Genetics 64, 18-23.

Otto, C., Yovkova, V., Aurich, A., Mauersberger, S., and Barth, G. (2012). Variation of the by-product spectrum during α-ketoglutaric acid production from raw glycerol by overexpression of fumarase and pyruvate carboxylase genes in *Yarrowia lipolytica*. Applied Microbiology and Biotechnology 95, 905-917.

Otto, S.P., and Whitton, J. (2000). Polyploid incidence and evolution. Annual Review of Genetics 34, 401-437.

Palmieri, L., Rottensteiner, H., Girzalsky, W., Scarcia, P., Palmieri, F., and Erdmann, R. (2001). Identification and functional reconstitution of the yeast peroxisomal adenine nucleotide transporter. EMBO J 20, 5049-5059.

Papanikolaou, S., Muniglia, L., Chevalot, I., Aggelis, G., and Marc, I. (2003). Accumulation of a Cocoa-Butter-Like Lipid by Yarrowia lipolytica Cultivated on Agro-Industrial Residues. Current Microbiology 46, 0124-0130.

Pâques, F., and Haber, J.E. (1999). Multiple Pathways of Recombination Induced by Double-Strand Breaks in *Saccharomyces cerevisiae*. Microbiology and Molecular Biology Reviews 63, 349-404.

Patrignani, F., Vannini, L., Gardini, F., Guerzoni, M.E., and Lanciotti, R. (2011). Variability of the lipolytic activity and volatile molecules production by a strain of *Yarrowia lipolytica* in pork fat and its dependence on environmental conditions. Meat Science 89, 21-26.

Payen, C., Koszul, R., Dujon, B., and Fischer, G. (2008). Segmental Duplications Arise from Pol32-Dependent Repair of Broken Forks through Two Alternative Replication-Based Mechanisms. PLoS Genet 4, e1000175.

Pellicer, J., Fay, M.F., and Leitch, I.J. (2010). The largest eukaryotic genome of them all? Botanical Journal of the Linnean Society 164, 10-15.

Peñalosa-Ruiz, G., Aranda, C., Ongay-Larios, L., Colon, M., Quezada, H., and Gonzalez, A. (2012). Paralogous ALT1 and ALT2 Retention and Diversification Have Generated Catalytically Active and Inactive Aminotransferases in *Saccharomyces cerevisiae*. PLoS ONE 7, e45702.

Péter, G., Dlauchy, D., Vasdinyei, R., Tornai-Lehoczki, J., and Deák, T. (2004). *Candida galli* sp. nov., a new yeast from poultry Antonie van Leeuwenhoek 86, 105-110-110.

Pignède, G., Wang, H., Fudalej, F., Gaillardin, C., Seman, M., and Nicaud, J.-M. (2000). Characterization of an Extracellular Lipase Encoded by LIP2 *in Yarrowia lipolytica*. Journal of Bacteriology 182, 2802-2810.

Porter, A. (2008). Preventing DNA over-replication: a Cdk perspective. Cell Division 3, 3.

Pronk, J.T., van der Linden-Beuman, A., Verduyn, C., Scheffers, W.A., and van Dijken, J.P. (1994). Propionate metabolism in *Saccharomyces cerevisiae*: implications for the metabolon hypothesis. Microbiology 140, 717-722.

Przybyla-Zawislak, B., Dennis, R.A., Zakharkin, S.O., and McCammon, M.T. (1998). Genes of succinyl-CoA ligase from *Saccharomyces cerevisiae*. European Journal of Biochemistry 258, 736-743.

Qian, W., Yang, J.-R., Pearson, N.M., Maclean, C., and Zhang, J. (2012). Balanced Codon Usage Optimizes Eukaryotic Translational Efficiency. PLoS Genet 8, e1002603.

Rachubinski, R.A., and Subramani, S. (1995). How proteins penetrate peroxisomes. Cell 83, 525-528.

Rasmussen, M.D., and Kellis, M. (2007). Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. Genome Research 17, 1932-1942.

Ratledge, C., and Wynn, J.P. (2002). The biochemistry and molecular biology of lipid accumulation in oleaginous microorganisms. Adv Appl Microbiol 51, 1-51.

Repetto, B., and Tzagoloff, A. (1991). In vivo assembly of yeast mitochondrial alpha-ketoglutarate dehydrogenase complex. Molecular and Cellular Biology 11, 3931-3939.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics *16*, 276-277.

Richard, G.-F., and Dujon, B. (2006). Molecular Evolution of Minisatellites in Hemiascomycetous Yeasts. Molecular Biology and Evolution 23, 189-202.

Rooney, A.P., and Ward, T.J. (2005). Evolution of a large ribosomal RNA multigene family in filamentous fungi: Birth and death of a concerted evolution paradigm. Proceedings of the National Academy of Sciences of the United States of America 102, 5084-5089.

Ross, J., Reid, G.A., and Dawes, I.W. (1988). The Nucleotide Sequence of the LPD1 Gene Encoding Lipoamide Dehydrogenase in *Saccharomyces cerevisiae*: Comparison between Eukaryotic and Prokaryotic Sequences for Related Enzymes and Identification of Potential Upstream Control Sites. Journal of General Microbiology 134, 1131-1139.

Sacerdot, C., Casaregola, S., Lafontaine, I., Tekaia, F., Dujon, B., and Ozier-Kalogeropoulos, O. (2008). Promiscuous DNA in the nuclear genomes of hemiascomycetous yeasts. FEMS Yeast Research 8, 846-857.

Sanderson, M.J. (1989). Confidence limits on phylogenies: the bootstrap revisited. Cladistics 5, 113-129.

Schacherer, J., Shapiro, J.A., Ruderfer, D.M., and Kruglyak, L. (2009). Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. Nature 458, 342-345.

Schrader, J., Etschmann, M.M.W., Sell, D., Hilmer, J.M., and Rabenhorst, J. (2004). Applied biocatalysis for the synthesis of natural flavour compounds - current industrial processes and future prospects. Biotechnology Letters 26, 463-472.

Schrag, J.D., Cygler, M., and Byron Rubin, E.A.D. (1997). Lipases and alpha/beta hydrolase fold. In Methods in Enzymology (Academic Press), pp. 85-107.

Sharp, P.M., Emery, L.R., and Zeng, K. (2010). Forces that influence the evolution of codon bias. Philosophical Transactions of the Royal Society B: Biological Sciences 365, 1203-1212.

Shockey, J., Chapital, D., Gidda, S., Mason, C., Davis, G., Klasson, K.T., Cao, H., Mullen, R., and Dyer, J. (2011). Expression of a lipid-inducible, self-regulating form of *Yarrowia*

lipolytica lipase LIP2 in Saccharomyces cerevisiae. Applied Microbiology and Biotechnology 92, 1207-1217.

Sonnhammer, E.L.L., and Koonin, E.V. (2002). Orthology, paralogy and proposed classification for paralog subtypes. Trends in genetics : TIG 18, 619-620.

Souciet, J.-L., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., Bon, E., Brottier, P., Casaregola, S., de Montigny, J., Dujon, B., et al. (2000). Genomic Exploration of the Hemiascomycetous Yeasts: 1. A set of yeast species for molecular evolution studies. FEBS Letters 487, 3-12.

Stadler, S.C., Polanetz, R., Meier, S., Mayerhofer, P.U., Herrmann, J.M., Anslinger, K., Roscher, A.A., R√∂schinger, W., and Holzinger, A. (2005). Mitochondrial targeting signals and mature peptides of 3-methylcrotonyl-CoA carboxylase. Biochemical and Biophysical Research Communications 334, 939-946.

Starmer, W.T., and Lachance, M.A. (2011). Yeast Ecology. In The Yeasts: a taxonomic study C.P. Kurtzman, J.W. Fell, and T. Boekhout, eds. (Elsevier), pp. 65-83.

Steensma, H.Y., Holterman, L., Dekker, I., van Sluis, C.A., and Wenzel, T.J. (1990). Molecular cloning of the gene for the E1α subunit of the pyruvate dehydrogenase complex from *Saccharomyces cerevisiae*. European Journal of Biochemistry 191, 769-774.

Steffan, J.S., and McAlister-Henn, L. (1992). Isolation and characterization of the yeast gene encoding the MDH3 isozyme of malate dehydrogenase. Journal of Biological Chemistry 267, 24708-24715.

Stoletzki, N., and Eyre-Walker, A. (2007). Synonymous Codon Usage in *Escherichia coli*: Selection for Translational Accuracy. Molecular Biology and Evolution 24, 374-381.

Suzzi, G., Lanorte, M.T., Galgano, F., Andrighetto, C., Lombardi, A., Lanciotti, R., and Guerzoni, M.E. (2001). Proteolytic, lipolytic and molecular characterisation of *Yarrowia lipolytica* isolated from cheese. International Journal of Food Microbiology 69, 69-77.

Swiegers, J.H., Dippenaar, N., Pretorius, I.S., and Bauer, F.F. (2001). Carnitine-dependent metabolic activities in *Saccharomyces cerevisiae*: three carnitine acetyltransferases are essential in a carnitine-dependent strain. Yeast 18, 585-595.

Szostak, J.W., and Wu, R. (1980). Unequal crossing over in the ribosomal DNA of *Saccharomyces cerevisiae*. Nature 284, 426-430.

Tabuchi, T., and Uchiyama, H. (1975). Methylcitrate Condensing and Methylisocitrate Cleaving Enzymes; Evidence for the Pathway of Oxidation of Propionyl-CoA to Pyruvate via C7-Tricarboxylic Acids. Agricultural and Biological Chemistry 39, 2035-2042.

Tai, M., and Stephanopoulos, G. (2013). Engineering the push and pull of lipid biosynthesis in oleaginous yeast *Yarrowia lipolytica* for biofuel production. Metabolic Engineering 15, 1-9.

Tehlivets, O., Scheuringer, K., and Kohlwein, S.D. (2007). Fatty acid synthesis and elongation in yeast. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids 1771, 255-270.

Tesler, G., Durand, D., Ma, J., Ratan, A., Zhang, L., Miller, W., and Haussler, D. (2007). A Heuristic Algorithm for Reconstructing Ancestral Gene Orders with Duplications. In Comparative Genomics (Springer Berlin Heidelberg), pp. 122-135.

The Génolevures Consortium, Souciet, J.-L., Dujon, B., Gaillardin, C., Johnston, M., Baret, P.V., Cliften, P., Sherman, D.J., Weissenbach, J., Westhof, E., et al. (2009). Comparative genomics of protoploid Saccharomycetaceae. Genome Research 19, 1696-1709.

Thevenieau, F., Le Dall, M.T., Nthangeni, B., Mauersberger, S., Marchal, R., and Nicaud, J.M. (2007). Characterization of *Yarrowia lipolytica* mutants affected in hydrophobic substrate utilization. Fungal Genetics and Biology 44, 531-542.

Thierry, A.s., Bouchier, C., Dujon, B., and Richard, G.-F. (2008). Megasatellites: a peculiar class of giant minisatellites in genes involved in cell adhesion and pathogenicity in Candida glabrata. Nucleic Acids Research 36, 5970-5982.

Thompson, L.M., Sutherland, P., Steffan, J.S., and McAlister-Henn, L. (1988). Gene sequence and primary structure of mitochondrial malate dehydrogenase from *Saccharomyces cerevisiae*. Biochemistry 27, 8393-8400.

Uchiyama, H., and Tabuchi, T. (1976). Properties of Methylcitrate Synthase from *Candida lipolytica*. Agricultural and Biological Chemistry 40, 1411-1418.

Uchiyama, H., Ando, M., Toyonaka, Y., and Tabuchi, T. (1982). Subcellular Localization of the Methylcitric-Acid-Cycle Enzymes in Propionate Metabolism of *Yarrowia Lipolytica*. European Journal of Biochemistry 125, 523-527.

van den Berg, M.A., de Jong-Gubbels, P., Kortland, C.J., van Dijken, J.P., Pronk, J.T., and Steensma, H.Y. (1996). The Two Acetyl-coenzyme A Synthetases of *Saccharomyces cerevisiae* Differ with Respect to Kinetic Properties and Transcriptional Regulation. Journal of Biological Chemistry 271, 28953-28959.

van Hoek, M.J.A., and Hogeweg, P. (2009). Metabolic Adaptation after Whole Genome Duplication. Molecular Biology and Evolution 26, 2441-2453.

van Roermund, C.W.T., Hettema, E.H., Kal, A.J., van den Berg, M., Tabak, H.F., and Wanders, R.J.A. (1998). Peroxisomal ß-oxidation of polyunsaturated fatty acids in *Saccharomyces cerevisiae*: isocitrate dehydrogenase provides NADPH for reduction of double bonds at even positions. EMBO J 17, 677-687.

van Roermund, C.W.T., Hettema, E.H., van den Berg, M., Tabak, H.F., and Wanders, R.J.A. (1999). Molecular characterization of carnitine-dependent transport of acetyl-CoA from peroxisomes to mitochondria in *Saccharomyces cerevisiae* and identification of a plasma membrane carnitine transporter, Agp2p. The EMBO Journal 18, 5843-5852.

van Roermund, C.W.T., Tabak, H.F., van den Berg, M., Wanders, R.J.A., and Hettema, E.H. (2000). Pex11p Plays a Primary Role in Medium-Chain Fatty Acid Oxidation, a Process That Affects Peroxisome Number and Size in *Saccharomyces cerevisiae*. The Journal of Cell Biology 150, 489-498.

Vanhanen, S., West, M., Kroon, J.T.M., Lindner, N., Casey, J., Cheng, Q., Elborough, K.M., and Slabas, A.R. (2000). A Consensus Sequence for Long-chain Fatty-acid Alcohol Oxidases from Candida Identifies a Family of Genes Involved in Lipid omega-Oxidation in Yeast with Homologues in Plants and Bacteria. Journal of Biological Chemistry 275, 4445-4452.

Verleur, N., Hettema, E.H., Van Roermund, C.W.T., Tabak, H.F., and Wanders, R.J.A. (1997). Transport of Activated Fatty Acids by the Peroxisomal ATP-binding-cassette Transporter Pxa2 in a Semi-Intact Yeast Cell System. European Journal of Biochemistry 249, 657-661.

Vorapreeda, T., Thammarongtham, C., Cheevadhanarak, S., and Laoteng, K. (2012). Alternative routes of acetyl-CoA synthesis identified by comparative genomic analysis: involvement in the lipid production of oleaginous yeast and fungi. Microbiology 158, 217-228.

Wakeley, J. (1996). The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. Trends in Ecology & Evolution 11, 158-162.

Walsh, J.B., and Stephan, W. (2001). Multigene Families: Evolution. In eLS (John Wiley & Sons, Ltd).

Walt, J.P., and Arx, J.A. (1980). The yeast genus *Yarrowia* gen. nov. Antonie van Leeuwenhoek 46, 517-521.

Wang, H., Dall, M.-T., Wach, Y., Laroche, C., Belin, J.-M., and Nicaud, J.-M. (1999). Cloning, sequencing, and characterization of five genes coding for Acyl-CoA oxidase isozymes in the yeast *Yarrowia lipolytica*. Cell Biochemistry and Biophysics 31, 165-174.

Wickerham, L.J., Kurtzman, C.P., and Herman, A.I. (1970). Sexual reproduction in *Candida lipolytica*. Science 167, 1141.

Wilks, H.M., Hart, K.W., Feeney, R., Dunn, C.R., Muirhead, H., Chia, W.N., Barstow, D.A., Atkinson, T., Clarke, A.R., and Holbrook, J.J. (1988). A specific, highly active malate dehydrogenase by redesign of a lactate dehydrogenase framework. Science (New York, NY) 242, 1541-1544.

Wolfe, K.H., and Shields, D.C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387, 708-713.

Wood, V., Gwilliam, R., Rajandream, M.A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., et al. (2002). The genome sequence of *Schizosaccharomyces pombe*. Nature 415, 871-880.

Yamagami, S., Iida, T., Nagata, Y., Ohta, A., and Takagi, M. (2001). Isolation and Characterization of Acetoacetyl-CoA Thiolase Gene Essential for n-Decane Assimilation in Yeast *Yarrowia lipolytica*. Biochemical and Biophysical Research Communications 282, 832-838.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution *24*, 1586-1591.

Yarrow, D. (1972). Four new combinations in yeasts. Antonie van Leeuwenhoek 38, 357-360.

Yovkova, V., Otto, C., Aurich, A., Mauersberger, S., and Barth, G. (2013). Engineering the α-ketoglutarate overproduction from raw glycerol by overexpression of the genes encoding NADP+-dependent isocitrate dehydrogenase and pyruvate carboxylase in *Yarrowia lipolytica*. Applied Microbiology and Biotechnology, 1-11.

Yuzbashev, T.V., Yuzbasheva, E.Y., Sobolevskaya, T.I., Laptev, I.A., Vybornaya, T.V., Larina, A.S., Matsui, K., Fukui, K., and Sineoky, S.P. (2010). Production of succinic acid at low pH by a recombinant strain of the aerobic yeast *Yarrowia lipolytica*. Biotechnology and Bioengineering 107, 673-682.

Zhang, H., Zhang, L., Chen, H., Chen, Y., Ratledge, C., Song, Y., and Chen, W. (2013). Regulatory properties of malic enzyme in the oleaginous yeast, *Yarrowia lipolytica*, and its non-involvement in lipid accumulation. Biotechnology Letters, 1-8.

Zhang, J. (2000). Rates of Conservative and Radical Nonsynonymous Nucleotide Substitutions in Mammalian Nuclear Genes. Journal of Molecular Evolution 50, 56-68.

Zhang, J., Rosenberg, H.F., and Nei, M. (1998). Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proceedings of the National Academy of Sciences 95, 3708-3713.

Zhang, L., Zhao, H., Xu, L., Liu, Y., and Yan, Y. (2010). Gene cloning, codon optimization and functional expression of *Yarrawia lipolytica* lipase Lip1. Wei Sheng Wu Xue Bao 50, 969-974.

Zhang, X., Li, M., Wei, D., and Xing, L. (2008). Identification and characterization of a novel yeast ω3-fatty acid desaturase acting on long-chain n-6 fatty acid substrates from Pichia pastoris. Yeast 25, 21-27.

Zhao, H., Xiao, X., Xu, L., Liu, Y., and Yan, Y. (2011a). Heterologous expression and characterization of Yarrowia lipolytica lipase 4 and lipase 5 in *Pichia pastoris*. Wei Sheng Wu Xue Bao 51, 1374-1381.

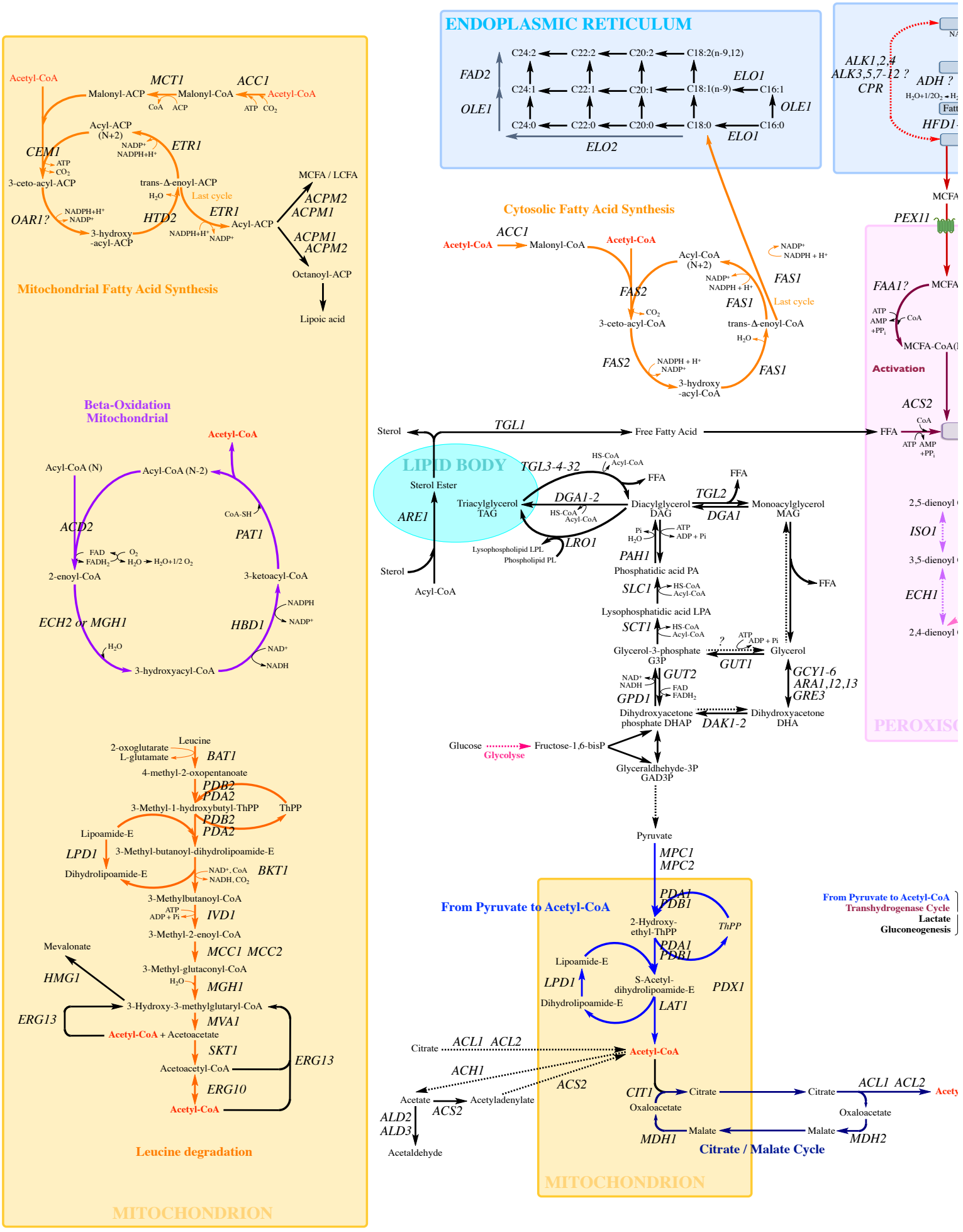
Zhao, H., Zheng, L., Wang, X., Liu, Y., Xu, L., and Yan, Y. (2011b). Cloning, expression and characterization of a new lipase from *Yarrowia lipolytica*. Biotechnology Letters 33, 2445-2452.

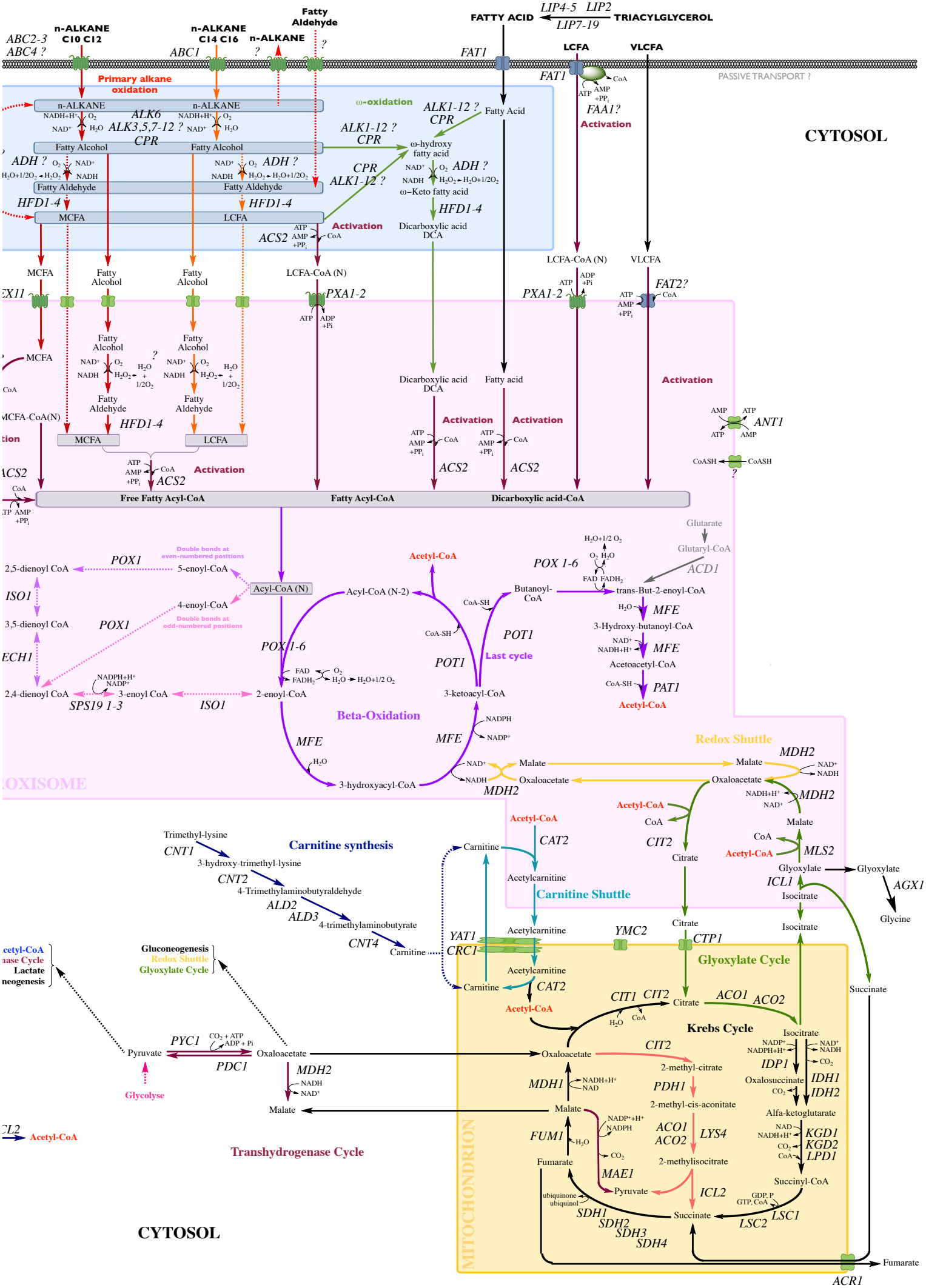
Zhou, J., Yin, X., Madzak, C., Du, G., and Chen, J. (2012). Enhanced α-ketoglutarate production in *Yarrowia lipolytica* WSH-Z06 by alteration of the acetyl-CoA metabolism. Journal of Biotechnology 161, 257-264.

Zimmer, E.A., Martin, S.L., Beverley, S.M., Kan, Y.W., and Wilson, A.C. (1980). Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. Proceedings of the National Academy of Sciences 77, 2158-2162.

Zou, Z., Tong, F., Faergeman, N.J., Børsting, C., Black, P.N., and DiRusso, C.C. (2003). Vectorial acylation in *Saccharomyces cerevisiae*. Fat1p and fatty acyl-CoA synthetase are interacting components of a fatty acid import complex. Journal of Biological Chemistry 278, 16414-16422.

ANNEXE 1


ABC2-3
ABC4 ?
ENDOPLASMIC RETICULUM
C24:2 C22:2 C20:2 C18:2(n-9,12)
C24:1 C22:1 C20:1 C18:1(n-9) C16:1
C24:0 C22:0 C20:0 C18:0 C16:0
FAD2
OLE1
ELO1
ELO2
ALK1,2,4
ALK3,5,7-12 ?
CPR
ADH ?
HFD1-
MCFA
PEX11
Cytosolic Fatty Acid Synthesis
ACC1
Acetyl-CoA
Malonyl-CoA
Acyl-CoA (N+2)
FAS2
FAS1
Last cycle
3-ceto-acyl-CoA
trans-Δ-enoyl-CoA
3-hydroxy -acyl-CoA
FAA1?
Activation
ACS2
FFA
LIPID BODY
Sterol
Sterol Ester
Triacylglycerol TAG
Diacylglycerol DAG
Monoacylglycerol MAG
TGL1
TGL3-4-32
DGA1-2
TGL2
DGA1
ARE1
LRO1
PAH1
Phosphatidic acid PA
SLC1
Lysophosphatidic acid LPA
SCT1
Glycerol-3-phosphate G3P
Glycerol
GUT1
GUT2
GPD1
GCY1-6
ARA1,12,13
GRE3
Dihydroxyacetone phosphate DHAP
Dihydroxyacetone DHA
DAK1-2
ISO1
ECH1
PEROXIS
Glucose
Glycolyse
Fructose-1,6-bisP
Glyceraldehyde-3P GAD3P
Pyruvate
MPC1
MPC2
From Pyruvate to Acetyl-CoA
PDA1
PDB1
2-Hydroxy-ethyl-ThPP
ThPP
Lipoamide-E
S-Acetyl-dihydrolipoamide-E
Dihydrolipoamide-E
LPD1
PDX1
LAT1
From Pyruvate to Acetyl-CoA
Transhydrogenase Cycle
Lactate
Gluconeogenesis
Citrate
ACL1 ACL2
ACH1
ACS2
Acetate
Acetyladenylate
ALD2
ALD3
Acetaldehyde
CIT1
Oxaloacetate
Malate
MDH1
MDH2
Citrate / Malate Cycle
MITOCHONDRION
Mitochondrial Fatty Acid Synthesis
MCT1
ACC1
Malonyl-ACP
Acyl-ACP (N+2)
CEM1
ETR1
3-ceto-acyl-ACP
trans-Δ-enoyl-ACP
OAR1?
HTD2
3-hydroxy -acyl-ACP
MCFA / LCFA
ACPM2
ACPM1
Octanoyl-ACP
Lipoic acid
Beta-Oxidation Mitochondrial
Acyl-CoA (N)
Acyl-CoA (N-2)
ACD2
PAT1
2-enoyl-CoA
3-ketoacyl-CoA
ECH2 or MGH1
HBD1
3-hydroxyacyl-CoA
Leucine degradation
Leucine
2-oxoglutarate
L-glutamate
BAT1
4-methyl-2-oxopentanoate
PDB2
PDA2
3-Methyl-1-hydroxybutyl-ThPP
ThPP
Lipoamide-E
LPD1
3-Methyl-butanoyl-dihydrolipoamide-E
Dihydrolipoamide-E
BKT1
3-Methylbutanoyl-CoA
IVD1
3-Methyl-2-enoyl-CoA
MCC1 MCC2
3-Methyl-glutaconyl-CoA
MGH1
Mevalonate
HMG1
3-Hydroxy-3-methylglutaryl-CoA
ERG13
MVA1
Acetoacetyl-CoA + Acetoacetate
SKT1
Acetoacetyl-CoA
ERG10
MITOCHONDRION


n-ALKANE C10 C12
n-ALKANE C14 C16
n-ALKANE
Fatty Aldehyde
ABC2-3
ABC4 ?
ABC1
FATTY ACID
LIP4-5 LIP2
LIP7-19
TRIACYLGLYCEROL
FAT1
LCFA
VLCFA
PASSIVE TRANSPORT ?
CYTOSOL
Primary alkane oxidation
ω-oxidation
ALK6
ALK3,5,7-12 ?
CPR
ALK1-12 ?
Fatty Alcohol
Fatty Aldehyde
ADH ?
HFD1-4
MCFA
LCFA
ω-hydroxy fatty acid
ω–Keto fatty acid
Dicarboxylic acid DCA
Fatty Acid
FAA1?
Activation
ACS2
LCFA-CoA (N)
PXA1-2
FAT2?
ANT1
Free Fatty Acyl-CoA
Fatty Acyl-CoA
Dicarboxylic acid-CoA
Glutarate
Glutaryl-CoA
ACD1
POX1
POX 1-6
ISO1
ECH1
SPS19 1-3
2,5-dienoyl CoA
5-enoyl-CoA
4-enoyl-CoA
3,5-dienoyl CoA
2,4-dienoyl CoA
3-enoyl CoA
2-enoyl-CoA
Acyl-CoA (N)
Acyl-CoA (N-2)
3-ketoacyl-CoA
3-hydroxyacyl-CoA
Butanoyl-CoA
trans-But-2-enoyl-CoA
3-Hydroxy-butanoyl-CoA
Acetoacetyl-CoA
MFE
POT1
PAT1
Last cycle
Beta-Oxidation
Acetyl-CoA
PEROXISOME
Redox Shuttle
MDH2
Malate
Oxaloacetate
Carnitine synthesis
Trimethyl-lysine
CNT1
3-hydroxy-trimethyl-lysine
CNT2
4-Trimethylaminobutyraldehyde
ALD2
ALD3
4-trimethylaminobutyrate
CNT4
Carnitine
CAT2
Acetylcarnitine
Carnitine Shuttle
YAT1
CRC1
YMC2
CTP1
CIT2
Citrate
MLS2
Glyoxylate
ICL1
Isocitrate
AGX1
Glycine
Glyoxylate Cycle
Krebs Cycle
CIT1
ACO1
ACO2
IDP1
IDH1
IDH2
Oxalosuccinate
Alfa-ketoglutarate
KGD1
KGD2
LPD1
Succinyl-CoA
LSC1
LSC2
Succinate
Fumarate
SDH1
SDH2
SDH3
SDH4
ACR1
MITOCHONDRION
PYC1
PDC1
Pyruvate
Oxaloacetate
Glycolyse
Gluconeogenesis
Redox Shuttle
Glyoxylate Cycle
Lactate neogenesis
Transhydrogenase Cycle
2-methyl-citrate
PDH1
2-methyl-cis-aconitate
ACO1 ACO2
LYS4
2-methylisocitrate
ICL2
MDH1
FUM1
MAE1
CYTOSOL

ANNEXE 2

Uniprot	locus_tag	Name	standard_name	EC number	Proposition for a new functional annotation	ancien /note de Génolevures	References	Evidence
Q6C670	YALI0E11979g	AAL1	YALI0E11979g1_1	EC 6.2.1.12	acyl/aryl-CoA ligase	weakly similar to uniprot|BAC71576 Streptomyces avermitilis SAV3864 Putative 4-coumarate:CoA ligase, hypothetical start		Electronic
Q6C8S6	YALI0D17314g	AAL10	YALI0D17314g1_1	EC 6.2.1.12	acyl/aryl-CoA ligase	weakly similar to uniprot|Q9K3W1 Streptomyces coelicolor 4-coumarate:CoA ligase		Electronic
Q6CH10	YALI0A14234g	AAL2	YALI0A14234g1_1	EC 6.2.1.12	acyl/aryl-CoA ligase	weakly similar to uniprot|AAN15615 Arabidopsis thaliana Putative 4-coumarate:CoA ligase 2 and to uniprot|P38137 Saccharomyces cerevisiae YBR222c PCS60 AMP-binding protein, peroxisomal, hypothetical start		Electronic
Q6C6V4	YALI0E05951g	AAL3	YALI0E05951g1_1	EC 6.2.1.12	acyl/aryl-CoA ligase	weakly similar to uniprot|Q9K3W1 Streptomyces coelicolor SCO4383 4-coumarate:CoA ligase, hypothetical start		Electronic
Q6C650	YALI0E12419g	AAL4	YALI0E12419g1_1	EC 6.2.1.12	acyl/aryl-CoA ligase	some similarities with uniprot|P38137 Saccharomyces cerevisiae YBR222c PCS60 AMP-binding protein, peroxisomal, hypothetical start		Electronic
Q6C2M7	YALI0F06556g	AAL5	YALI0F06556g1_1	EC 6.2.1.12	acyl/aryl-CoA ligase	weakly similar to uniprot|P38137 Saccharomyces cerevisiae YBR222C PCS60 Peroxisomal AMP-binding protein, putative acyl-CoA synthetase, hypothetical start		Electronic
Q6CCW9	YALI0C05885g	AAL6	YALI0C05885g1_1	EC 6.2.1.12	acyl/aryl-CoA ligase	similar to uniprot|Q8S564 Glycine max 4-coumarate:coenzyme A ligase (EC 6.2.1.12), hypothetical start		Electronic
Q6C577	YALI0E20405g	AAL7	YALI0E20405g1_1	EC 6.2.1.12	acyl/aryl-CoA ligase	weakly similar to uniprot|O48868 Populus balsamifera 4-coumarate:CoA ligase 2, hypothetical start		Electronic
Q6CFE4	YALI0B07755g	AAL8	YALI0B07755g1_1	EC 6.2.1.12	acyl/aryl-CoA ligase	weakly similar to uniprot|Q7SDW1 Neurospora crassa NCU03295.1 related to 4-coumarate-CoA ligase, hypothetical start		Electronic
Q6CGX7	YALI0A15103g	AAL9	YALI0A15103g1_1	EC 6.2.1.12	acyl/aryl-CoA ligase	weakly similar to uniprot|Q7S557 Neurospora crassa NCU03295.1 hypothetical protein, hypothetical start		Electronic
Q6C5V9	YALI0E14729g	ABC1	YALI0E14729g1_1		Plasma membrane ATP-binding cassette (ABC) transporter	similar to uniprot|P33302 Saccharomyces cerevisiae YOR153w PDR5 (ohnolog of YDR406W) Plasma membrane ATP-binding cassette (ABC) transporter, multidrug transporter actively regulated by Pdr1p	Thevenieau et al. (2007) Fungal Genet Biol 44:531–542; Chen B, Ling H, Chang MW (2013) Biotechnol Biofuels 6(1):21	Gene disruption; Expression in S. cerevisiae, substrate specificity
Q6CBB8	YALI0C20265g	ABC2	YALI0C20265g1_1		Plasma membrane ATP-binding cassette (ABC) transporter	uniprot|Q6CBB8 Yarrowia lipolytica YALI0C20265g YlABC2	Chen B, Ling H, Chang MW (2013) Biotechnol Biofuels 6(1):21	Expression in S. cerevisiae, substrate specificity
Q6CFY4	YALI0B02544g	ABC3	YALI0B02544g1_1		Plasma membrane ATP-binding cassette (ABC) transporter	similar to uniprot|Q04182 Saccharomyces cerevisiae YDR406w PDR15 ATP-binding cassette transporter family member	Chen B, Ling H, Chang MW (2013) Biotechnol Biofuels 6(1):21	Expression in S. cerevisiae, substrate specificity
Q6CET9	YALI0B12980g	ABC4	YALI0B12980g1_1		Plasma membrane ATP-binding cassette (ABC) transporter	similar to uniprot|Q96VK4 Emericella nidulans ABC transporter protein	Chen B, Ling H, Chang MW (2013) Biotechnol Biofuels 6(1):21	Expression in S. cerevisiae, substrate specificity
Q6C1A3	YALI0F17996g	ABC5lf	YALI0F17996g2_1		Plasma membrane ATP-binding cassette (ABC) transporter	similar to uniprot|P32568 Saccharomyces cerevisiae YDR011w SNQ2 multidrug resistance protein, hypothetical start		
Q6C1A3	YALI0F17996g	ABC5sf	YALI0F17996g1_1		Plasma membrane ATP-binding cassette (ABC) transporter	similar to uniprot|P32568 Saccharomyces cerevisiae YDR011w SNQ2 multidrug resistance protein, hypothetical start		
Q6CC91	YALI0C11407g	ACC1	YALI0C11407g1_1		Acetyl-CoA-carboxylase	similar to uniprot|Q00955 Saccharomyces cerevisiae YNR016C ACC1 (ohnolog of YMR207C) Acetyl-CoA carboxylase		Electronic
Q6C0L1	YALI0F23749g	ACD1	YALI0F23749g1_1		Glutaryl-CoA dehydrogenase	highly similar to uniprot|Q7S579 Neurospora crassa NCU02291.1 probable Glutaryl-CoA dehydrogenase		
Q6C0L1	YALI0D15708g	ACD2	YALI0D15708g1_1		Acyl CoA dehydrogenase	similar to uniprot|P45954 Homo sapiens Acyl-CoA dehydrogenase short/branched chain specific mitochondrial precursor		
Q6CBP3	YALI0C16797g	ACD3	YALI0C16797g1_1		Acyl CoA dehydrogenase	similar to uniprot|Q96VP9 Glomus intraradices Probable acyl-CoA dehydrogenase		
Q6CFQ1	YALI0B04906g	ACD4	YALI0B04906g1_1		Acyl CoA dehydrogenase	similar to uniprot| G9P1F6 Trichoderma atroviride very-long-chain acyl-CoA dehydrogenase		
Q6C3Z9	YALI0E30965g	ACH1	YALI0E30965g1_1		CoA transferase for CoASH transfer from succinyl-CoA to acetate	similar to uniprot|P32316 Saccharomyces cerevisiae YBL015W ACH1 Acetyl-coA hydrolase		
Q6C3H5	YALI0E34793g	ACL1	YALI0E34793g1_1	EC 2.3.3.8	ATP citrate lyase	highly similar to uniprot|Q8X097 Neurospora crassa Probable ATP citrate lyase subunit 1	Liu XY, Chi Z, Liu GL, Madzak C, Chi ZM (2013) Mar Biotechnol (NY) 15(1):26-36	Gene disruption, citric acid accumulation
Q6C7Y1	YALI0D24431g	ACL2	YALI0D24431g1_1	EC 2.3.3.8	ATP citrate lyase	similar to uniprot|Q8X096 Neurospora crassa Probable ATP citrate lyase subunit 2	Liu XY, Chi Z, Liu GL, Madzak C, Chi ZM (2013) Mar Biotechnol (NY) 15(1):26-36	Gene disruption, citric acid accumulation
Q6C9P6	YALI0D09361g	ACO1	YALI0D09361g1_1		Aconitase	highly similar to uniprot|P19414 Saccharomyces cerevisiae YLR304C ACO1 Aconitase	Holz M, Förster A, Mauersberger S, Barth G (2009) Appl Microbiol Biotechnol. 81: 1087-1096	Cloning, overexpression and metabolic effects
Q6C5V2	YALI0E14949g	ACO2	YALI0E14949g1_1		Aconitase	highly similar to uniprot|P39533 Saccharomyces cerevisiae YJL200c strong similarity to aconitate hydratase	Uchiyama H, Ando M, Toyonaka Y & Tabuchi T (1982) Eur J Biochem 125, 523–527	Subcellular localisation of enzyme activities
Q6C926	YALI0D14850g	ACPM1	YALI0D14850g1_1		mitochonrial acyl-carrier protein	similar to uniprot|P32463 Saccharomyces cerevisiae YKL192c ACP1 mitochonrial acyl-carrier protein	Dobrynin K, Abdrakhmanova A, Richers S, Hunte C, Kerscher S, Brandt U (2010) Biochim Biophys Acta 1797(2):152-9	Deletion, subcellular localization, Western blotting
Q6C7X2	YALI0D24629g	ACPM2	YALI0D24629g1_1		mitochonrial acyl-carrier protein	weakly similar to uniprot|P32463 Saccharomyces cerevisiae YKL192c ACP1 mitochonrial acyl-carrier protein	Dobrynin K, Abdrakhmanova A, Richers S, Hunte C, Kerscher S, Brandt U (2010) Biochim Biophys Acta 1797(2):152-9	Deletion, subcellular localization, Western blotting
Q6C3I0	YALI0E34672g	ACR1	YALI0E34672g1_1		succinate-fumarate transporter	similar to uniprot|P33303 Saccharomyces cerevisiae YJR095w ACR1 succinate-fumarate transporter P33.2.f7.1		
Q6C2Q5	YALI0F05962g	ACS2	YALI0F05962g1_1		Acetyl-coA synthetase	similar to uniprot|P52910 Saccharomyces cerevisiae YLR153c ACS2 Acetyl-coA synthetase isoform which, along with Acs1p, is the nuclear source of acetyl-coA for histone acetylation		
Q9UW08	YALI0D25630g	ADH1	YALI0D25630g1_1		Alcohol dehydrogenase	uniprot|Q9UW08 Yarrowia lipolytica Alcohol dehydrogenase 1, identified start	Barth G and Künkel W (1979) Z. Allg. Mikrobiol. 19:381-390	
Q9UW07	YALI0E17787g	ADH2	YALI0E17787g1_1		Alcohol dehydrogenase II	uniprot|Q9UW07 Yarrowia lipolytica YALI0E17787g Alcohol dehydrogenase 2	Barth G and Künkel W (1979) Z. Allg. Mikrobiol. 19:381-390	
Q9UW06	YALI0A16379g	ADH3	YALI0A16379g1_1		Alcohol dehydrogenase II	uniprot|Q9UW06 Yarrowia lipolytica YALI0A16379g Alcohol dehydrogenase 3	Barth G and Künkel W (1979) Z. Allg. Mikrobiol. 19:381-390	
Q6CGX5	YALI0A15147g	ADH4	YALI0A15147g1_1		Alcohol dehydrogenase III	similar to uniprot|Q9UW06 Yarrowia lipolytica Alcohol dehydrogenase 3		
Q6C6P0	YALI0E07766g	ADH5	YALI0E07766g1_1		Alcohol dehydrogenase I	similar to uniprot|P07246 Saccharomyces cerevisiae YMR083w ADH3 alcohol dehydrogenase III		
Q6C5N1	YALI0E16643g	AGX1	YALI0E16643g1_1		alanine-glyoxylate aminotransferase	similar to uniprot|P43567 Saccharomyces cerevisiae YFL030w		
Q6C2W9	YALI0F04444g	ALD2	YALI0F04444g1_1		Cytoplasmic aldehyde dehydrogenase	similar to uniprot|Q59N06 Candida albicans Trimethylaminobutyraldehyde dehydrogenase, the third enzyme of the carnitine biosynthesis pathway		

Q6C9V7	YALI0D07942g	ALD3	YALI0D07942g1_1		Cytoplasmic aldehyde dehydrogenase	similar to uniprot|Q59N06 Candida albicans Trimethylaminobutyraldehyde dehydrogenase, the third enzyme of the carnitine biosynthesis pathway		
O74127	YALI0E25982g	ALK1	YALI0E25982g1_1		n-alkane- inducible cytochrome P450	uniprot|O74127 Yarrowia lipolytica YALI0E25982g ALK1 n-alkane- inducible cytochrome P450	Iida T, Sumita T, Ohta A, Takagi M (2000) Yeast 16(12):1077-87	Cloning, sequencing, mutation, phenotype
Q6CDW4	YALI0B20702g	ALK10	YALI0B20702g1_1		n-alkane- inducible cytochrome P451	similar to uniprot|O74128 Yarrowia lipolytica ALK2	Takai H et al. (2012) Fungal Genet Biol 49:58-64	Cloning, mutagenesis, multiple mutants, phenotype
Q6CCE5	YALI0C10054g	ALK11	YALI0C10054g1_1		n-alkane- inducible cytochrome P452	similar to uniprot|O74128 Yarrowia lipolytica ALK2	Takai H et al. (2012) Fungal Genet Biol 49:58-64	Cloning, mutagenesis, multiple mutants, phenotype
Q6CGD9	YALI0A20130g	ALK12	YALI0A20130g1_1		n-alkane- inducible cytochrome P453	similar to uniprot|O74129 Yarrowia lipolytica ALK3 cytochrome P450	Takai H et al. (2012) Fungal Genet Biol 49:58-64	Cloning, sequencing, mutation, phenotype
O74128	YALI0F01320g	ALK2	YALI0F01320g1_1		n-alkane- inducible cytochrome P454	uniprot|O74128 Yarrowia lipolytica YALI0F01320g ALK2 (CytP450 familly)	Iida T, Sumita T, Ohta A, Takagi M (2000) Yeast 16(12):1077-87	Cloning, sequencing, mutation, phenotype
O74129	YALI0E23474g	ALK3	YALI0E23474g1_1		n-alkane- inducible cytochrome P455	uniprot|O74129 Yarrowia lipolytica YALI0E23474g alkane utilisation CYP52 family member	Hanley K et al. (2003) Assay Drug Dev Technol. 1(1 Pt 2):147-60.	Expression in plants, substrate specificity
O74130	YALI0B13816g	ALK4	YALI0B13816g1_1		n-alkane- inducible cytochrome P456	uniprot|O74130 Yarrowia lipolytica YALI0B13816g ALK4 CYP52 family member	Iida T, Sumita T, Ohta A, Takagi M (2000) Yeast 16(12):1077-87	Cloning, sequencing, mutation, phenotype
O74131	YALI0B13838g	ALK5	YALI0B13838g1_1		n-alkane- inducible cytochrome P457	uniprot|O74131 Yarrowia lipolytica YALI0B13838g ALK5 CYP52 family member	Hanley K et al. (2003) Assay Drug Dev Technol. 1(1 Pt 2):147-60.	Expression in plants, substrate specificity
O74132	YALI0B01848g	ALK6	YALI0B01848g1_1		n-alkane- inducible cytochrome P458	uniprot|O74132 Yarrowia lipolytica YALI0B01848g ALK6 CYP52 family member	Iida T, Sumita T, Ohta A, Takagi M (2000) Yeast 16(12):1077-87	Cloning, sequencing, mutation, phenotype
O74133	YALI0A15488g	ALK7	YALI0A15488g1_1		n-alkane- inducible cytochrome P459	uniprot|O74133 Yarrowia lipolytica YALI0A15488g ALK7 cytochrome P450 CYP52 family member	Hanley K et al. (2003) Assay Drug Dev Technol. 1(1 Pt 2):147-60.	Expression in plants, substrate specificity
O74134	YALI0C12122g	ALK8	YALI0C12122g1_1		n-alkane- inducible cytochrome P460	uniprot|O74134 Yarrowia lipolytica YALI0C12122g ALK8 cytochrome P450	Takai H et al. (2012) Fungal Genet Biol 49:58-64	Cloning, mutagenesis, multiple mutants, phenotype, ALK1 induction
Q6CFK2	YALI0B06248g	ALK9	YALI0B06248g1_1		n-alkane- inducible cytochrome P461	similar to uniprot|O74127 Yarrowia lipolytica ALK1	Takai H et al. (2012) Fungal Genet Biol 49:58-64	Cloning, mutagenesis, multiple mutants, phenotype, ALK1 induction
Q6C780	YALI0E03058g	ANT1	YALI0E03058g1_1		Peroxisomal ATP/AMP transporter	similar to uniprot|Q06497 Saccharomyces cerevisiae YPR128C ANT1 Peroxisomal adenine nucleotide transporter involved in beta-oxidation of medium-chain fatty acid	Thevenieau et al. (2007) Fungal Genet Biol 44:531–542; Chen B, Ling H, Chang MW (2013) Biotechnol Biofuels 6(1):21	Gene disruption, overexpression
Q6CCV8	YALI0C06127g	ANT2	YALI0C06127g1_1		Putative peroxisomal ATP/AMP transporter	weakly similar to uniprot|Q06497 Saccharomyces cerevisiae YPR128c ANT1 Peroxisomal adenine nucleotide transporter		
Q6C2L0	YALI0F06974g	ARA1	YALI0F06974g1_1	EC 1.1.1.-	protein of aldo/keto reductase family	similar to uniprot|P38115 Saccharomyces cerevisiae YBR149w ARA1 D-arabinose dehydrogenase large subunit	Dulermo T et al. (2011) Metabolic Engineering 13, 482-491	
Q6CC16	YALI0C13508g	ARA12	YALI0C13508g1_1	EC 1.1.1.-	protein of aldo/keto reductase family	similar to uniprot|P14065 Saccharomyces cerevisiae YOR120w GCY1 galactose-induced protein of aldo/keto reductase family P6.5.f6.1 (EC 1.1.1.-)	Dulermo T et al. (2011) Metabolic Engineering 13, 482-491	
Q6CCI5	YALI0C09119g	ARA13	YALI0C09119g1_1	EC 1.1.1.-	protein of aldo/keto reductase family	similar to uniprot|P14065 Saccharomyces cerevisiae YOR120W GCY protein	Dulermo T et al. (2011) Metabolic Engineering 13, 482-491	
Q6C2M6	YALI0F06578g	ARE1	YALI0F06578g1_1		acyl-CoA sterol acyltransferase	similar to uniprot|P53629 Saccharomyces cerevisiae YNR019w ARE2 acyl-CoA sterol acyltransferase		
Q6CAN4	YALI0D01265g	BAT1	YALI0D01265g1_1	EC 2.6.1.42	Branched-chain Amino acid Transaminase, mitochondrial	similar to uniprot|P38891 Saccharomyces cerevisiae YHR208w BAT1 branched chain amino acid aminotransferase mitochondrial precursor	Bondar DC et al. (2005) AEM 71: 4585-4591	
Q6C120	YALI0F19910g	BAT2	YALI0F19910g1_1		Branched-chain Amino acid Transaminase, cytoplasmic	similar to uniprot|P47176 Saccharomyces cerevisiae YJR148w TWT2 branched chain amino acid aminotransferase cytosolic	Bondar DC et al. (2005) AEM 71: 4585-4591	
Q6C806	YALI0D23815g	BKT1	YALI0D23815g1_1	EC 2.3.1.168	Branched-chain keto acid dehydrogenase E1	similar to tuniprot|P53395 Mus musculus Lipoamide acyltransferase component of branched-chain alpha-keto acid dehydrogenase	Vorapreeda et al. (2012) Microbiology, 158, 217–228	
Q6CF46	YALI0B10340g	CAT2	YALI0B10340g1_1		Peroxisomal carnitine acetyl transferase, cytosolic	similar to uniprot|O42620 Magnaporthe grisea Carnitine acetyl transferase		
Q6BZV1	YALI0F30679g	CEM1	YALI0F30679g1_1		3-oxoacyl-[acyl-carrier protein] synthase	similar to uniprot|P39525 Saccharomyces cerevisiae YER061c CEM1 3-oxoacyl-[acyl-carrier protein] synthase		
Q6C7I2	YALI0E00638g	CIT1	YALI0E00638g1_1		Mitochondrial 2-methyl citrate synthase	similar to uniprot|Q9TEM3 Emericella nidulans MCSA similar to uniprot|Q9TEM3 Emericella nidulans MCSA Methylcitrate synthase precursor	Uchiyama H, Ando M, Toyonaka Y & Tabuchi T (1982) Eur J Biochem 125, 523–527	
Q6C793	YALI0E02684g	CIT2	YALI0E02684g1_1		Mitochondrial citrate synthase	highly similar to uniprot|P00890 Saccharomyces cerevisiae YNR001c CIT1 (ohnolog of YCR005C) Citrate synthase		
Q6CCC7	YALI0C10604g	CNT1	YALI0C10604g1_1		Trimethyllysine dioxygenase	similar to uniprot|Q5AG63 Candida albicans Trimethyllysine dioxygenase, the first enzyme in the carnitine biosynthesis pathway		
Q6CCC7	YALI0A21417g	CNT2	YALI0A21417g1_1		Hydroxytrimethyllysine aldolase	similar to uniprot|Q59N07 Candida albicans Hydroxytrimethyllysine aldolase, the second enzyme in the carnitine biosynthesis pathway		
Q6CCC7	YALI0F16357g	CNT4	YALI0F16357g1_1		Butyrobetaine dioxygenase	similar to uniprot|Q5A0G4 Candida albicans Butyrobetaine dioxygenase, the fourth enzyme of the carnitine biosynthesis pathway		
Q6L8F1	YALI0D04422g	CPR1	YALI0D04422g1_1		NADPH-cytochrome P450 reductase	uniprot|Q6L8F1 Yarrowia lipolytica YALI0D04422g YlCPR1 NADPH-cytochrome P450 reductase	Novikova et al. (2009) Biochemistry (Mosc)74: 1482-1504.	
Q6CDA3	YALI0C02431g	CRC1	YALI0C02431g1_1		Mitochondrial carnitine carrier protein	similar to uniprot|Q9Y7G4 Emericella nidulans Carnitine/acyl carnitine carrier		
Q6C564	YALI0E20691g	DAK1	YALI0E20691g1_1		dihydroxyacetone kinase	similar to uniprot|P54838 Saccharomyces cerevisiae YML070w dihydroxyacetone kinase	Dulermo T et al. (2011) Metabolic Engineering 13:482-491	
Q6C2B3	YALI0F09273g	DAK2	YALI0F09273g1_1		dihydroxyacetone kinase	similar to uniprot|P43550 Saccharomyces cerevisiae YFL053w DAK2 Dihydroxyacetone kinase 2 or uniprot|P54838 Saccharomyces cerevisiae YML070w DAK1 dihydroxyacetone kinase and DEHA0A03971g Debaryomyces hansenii	Dulermo T et al. (2011) Metabolic Engineering 13:482-491	
Q6C394	YALI0F01606g	DAK3	YALI0F01606g1_1		dihydroxyacetone kinase	similar to uniprot|O13902 Schizosaccharomyces pombe Dihydroxyacetone kinase 1 DAK1 (DHA kinase 1)		
Q6C3R2	YALI0E32769g	DGA1	YALI0E32769g1_1	EC 2.3.1.20	acyl-CoA:diacylglycerol acyltransferase	some similarities with uniprot|Q08650 Saccharomyces cerevisiae YOR245C DGA1 acyl-CoA:diacylglycerol acyltransferase	Blazeck et al. (2014) Nature Communications 5:3131 DOI: 10.1038/ncomms4131	Overexpression
Q6C9V5	YALI0D07986g	DGA2	YALI0D07986g1_1	EC 2.3.1.20	acyl-CoA:diacylglycerol acyltransferase	weakly similar to uniprot|Q9FUL6 Perilla frutescens Diacylglycerol acyltransferase	Blazeck et al. (2014) Nature Communications 5:3131 DOI: 10.1038/ncomms4131	Overexpression

Q6CCU5	YALI0C06446g	DLD2	YALI0C06446g1_1		Lactate dehydrogenase	similar to uniprot|P46681 Saccharomyces cerevisiae YDL178W DLD2 D-lactate dehydrogenase		
Q6CHL1	YALI0A07733g	ECH1	YALI0A07733g1_1	EC 4.2.1.17	Enoyl-CoA hydratase	similar to uniprot|Q5BAA1 Aspergillus nidulans AN2529.2 enoyl-CoA hydratase/isomerase family protein		
Q6CF43	YALI0B10406g	ECH2	YALI0B10406g1_1	EC 4.2.1.17	Enoyl-CoA hydratase	similar to uniprot|A2QCA2 Aspergillus niger enoyl-CoA hydratase		
Q6CA37	YALI0D06215g	EHD3	YALI0D06215g1_1	EC 4.2.1.17	Enoyl-CoA hydratase	similar to uniprot|Q9C251 Neurospora crassa Related to enoyl-CoA-hydratase		
Q6CDY7	YALI0B20196g	ELO1	YALI0B20196g1_1		fatty acid elongase required for sphingolipid formation	similar to uniprot|P25358 Saccharomyces cerevisiae YCR034w GNS1 fatty acid elongase required for sphingolipid formation	Dittrich F et al. (1998) Eur. J. Biochem. 252:477-485	
Q6C2L8	YALI0F06754g	ELO2	YALI0F06754g1_1		fatty acid elongase required for sphingolipid formation	similar to ca|CA6061|IPF8302 Candida albicans	Dittrich F et al. (1998) Eur. J. Biochem. 252:477-485	
Q6CFB6	YALI0B08536g	ERG10	YALI0B08536g1_1		Acetyl-CoA C-acetyltransferase (acetoacetyl-CoA thiolase)	similar to uniprot|Q04677 Candida tropicalis Acetyl-CoA acetyltransferase I		
Q6BZW0	YALI0F30481g	ERG13	YALI0F30481g1_1		3-hydroxy-3-methylglutaryl-CoA (HMG-CoA) synthase	similar to uniprot|P54839 Saccharomyces cerevisiae YML126c ERG13 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA) synthase		
Q6CBE4	YALI0C19624g	ETR1	YALI0C19624g1_1		2-enoyl thioester reductase, member of the medium chain dehydrogenase/reductase family	similar to uniprot|P38071 Saccharomyces cerevisiae YBR026c (MRF1) mitochondrial respiratory function protein		
Q6C8Q3	YALI0D17864g	FAA1	YALI0D17864g1_1		Long chain fatty acyl-CoA synthetase	similar to uniprot|P30624 Saccharomyces cerevisiae YOR317w FAA1 (ohnolog of YMR246W) Long chain fatty acyl-CoA synthetase with a preference for C12:0-C16:0 fatty acids		
Q6CF55	YALI0B10153g	FAD2	YALI0B10153g1_1		oleate delta-12 desaturase	similar to uniprot|Q9HF05 Emericella nidulans Oleate delta-12 desaturase	Ferrante G et al. (1983) CAN. J. BIBCHEM. CELL BIOL. VOL. 61; Beopoulos et al. (2014) Appl Microbiol Biotechnol 98(1):251-62. doi: 10.1007/s00253-013-5295-x	
P34229	YALI0B15059g	FAS1	YALI0B15059g1_1	EC 2.3.1.86	Fatty acid synthase subunit beta	uniprot|P34229 Yarrowia lipolytica Fatty acid synthase subunit beta (EC 2.3.1.86)		
Q6CE16	YALI0B19382g	FAS2	YALI0B19382g1_1	EC 2.3.1.86	Fatty acid synthase subunit alpha	similar to uniprot|P43098 Candida albicans Fatty acid synthase subunit alpha (EC 2.3.1.86)		
Q6C5Q8	YALI0E16016g	FAT1	YALI0E16016g1_1		Very long chain fatty acyl-CoA synthetase and long chain fatty acid transporter	similar to uniprot|P38225 Saccharomyces cerevisiae YBR041w FAT1 Fatty acid transporter and very long-chain fatty acyl-CoA synthetase, may form a complex with Faa1p or Faa4p that imports and activates exogenous fatty acids		
Q6C634	YALI0E12859g	FAT2	YALI0E12859g1_1		Peroxisomal protein that binds AMP and mRNA	similar to uniprot|P38137 Saccharomyces cerevisiae YBR222c PCS60 AMP-binding protein peroxisomal-coenzyme A synthetase		
Q6CFN2	YALI0B05456g	FAT3	YALI0B05456g1_1		Peroxisomal protein that binds AMP and mRNA	weakly similar to uniprot|P38137 Saccharomyces cerevisiae YBR222c PCS60 AMP binding protein peroxisomal		
Q6CCH6	YALI0C09284g	FAT4	YALI0C09284g1_1		Peroxisomal protein that binds AMP and mRNA	some similarities with uniprot|P38137 Saccharomyces cerevisiae YBR222c PCS60 AMP-binding protein peroxisomal		
Q6CCT2	YALI0C06776g	FUM1	YALI0C06776g1_1	EC 4.2.1.2	Fumarate hydratase mitochondrial precursor (Fumarase)	highly similar to uniprot|P55250 Rhizopus oryzae Fumarate hydratase mitochondrial precursor (EC 4.2.1.2) (Fumarase)	Morín M, Monteoliva L, Insenser M, Gil C, Domínguez A (2007) J Mass Spectrom 42(11):1453-62	Identified by 2D-gel analysis as associated to hyphae formation
Q6CAC3	YALI0D04092g	GCY1	YALI0D04092g1_1		galactose-induced protein of aldo/keto reductase family	similar to uniprot|P14065 Saccharomyces cerevisiae YOR120w GCY1 galactose-induced protein of aldo/keto reductase family	Dulermo T et al. (2011) Metabolic Engineering 13:482-491	
Q6CFG7	YALI0B07117g	GCY12	YALI0B07117g1_1		galactose-induced protein of aldo/keto reductase family	similar to uniprot|P14065 Saccharomyces cerevisiae YOR120w GCY1 galactose-induced protein of aldo/keto reductase family	Dulermo T et al. (2011) Metabolic Engineering 13:482-491	
Q6CFG7	YALI0A15906g	GCY13	YALI0A15906g1_1		galactose-induced protein of aldo/keto reductase family	similar to uniprot|P14065 Saccharomyces cerevisiae YOR120w GCY1 galactose-induced protein of aldo/keto reductase family	Dulermo T et al. (2011) Metabolic Engineering 13:482-491	
Q6CFG7	YALI0E18348g	GCY14	YALI0E18348g1_1		galactose-induced protein of aldo/keto reductase family	similar to uniprot|P14065 Saccharomyces cerevisiae YOR120w GCY1 galactose-induced protein of aldo/keto reductase family	Dulermo T et al. (2011) Metabolic Engineering 13:482-491	
Q6CFG7	YALI0F18590g	GCY15	YALI0F18590g1_1		galactose-induced protein of aldo/keto reductase family	similar to uniprot|P14065 Saccharomyces cerevisiae YOR120w GCY1 galactose-induced protein of aldo/keto reductase or uniprot|Q12458 Saccharomyces cerevisiae YDR368w YPR1 strong similarity to members of the aldo/keto reductase family	Dulermo T et al. (2011) Metabolic Engineering 13:482-491	
Q6CFG7	YALI0B21780g	GCY16	YALI0B21780g1_1		galactose-induced protein of aldo/keto reductase family	similar to uniprot|P14065 Saccharomyces cerevisiae YOR120w GCY1 galactose-induced protein of aldo/keto reductase family	Dulermo T et al. (2011) Metabolic Engineering 13:482-491	
Q9UVF4	YALI0B02948g	GPD1	YALI0B02948g1_1		Glycerol-3- phosphate dehydrogenase [NAD]	uniprot|Q9UVF4 Yarrowia lipolytica Glycerol-3- phosphate dehydrogenase [NAD]	Dulermo T et al. (2011) Metabolic Engineering 13:482-491	
Q6C9X2	YALI0D07634g	GRE3	YALI0D07634g1_1		aldose reductase	similar to uniprot|P38715 Saccharomyces cerevisiae YHR104w GRE3 aldose reductase	Dulermo T et al. (2011) Metabolic Engineering 13:482-491	
Q6C3E1	YALI0F00484g	GUT1	YALI0F00484g1_1	EC 2.7.1.30	Glycerol kinase	similar to uniprot|Q63060 Rattus norvegicus Glycerol kinase (EC 2.7.1.30)	Dulermo T et al. (2011) Metabolic Engineering 13:482-491	
Q6CEQ0	YALI0B13970g	GUT2	YALI0B13970g1_1		glycerol-3-phosphate dehydrogenase mitochondrial	similar to uniprot|P32191 Saccharomyces cerevisiae YIL155c GUT2 glycerol-3-phosphate dehydrogenase mitochondrial	Dulermo T et al. (2011) Metabolic Engineering 13:482-491	
Q6C0L0	YALI0F23793g	HFD1	YALI0F23793g1_1		Putative fatty aldehyde dehydrogenase	similar to uniprot|Q04458 Saccharomyces cerevisiae YMR110c HFD1 Putative fatty aldehyde dehydrogenase, located in the mitochondrial outer membrane and also in lipid particles		
Q6C5T1	YALI0E15400g	HFD2lf	YALI0E15400g1_1		Putative fatty aldehyde dehydrogenase	similar to uniprot|Q04458 Saccharomyces cerevisiae YMR110c HFD1 Putative fatty aldehyde dehydrogenase, located in the mitochondrial outer membrane and also in lipid particles		
Q6C5T1	YALI0E15400g	HFD2sf	YALI0E15400g2_1		Putative fatty aldehyde dehydrogenase	similar to uniprot|Q04458 Saccharomyces cerevisiae YMR110c HFD1 Putative fatty aldehyde dehydrogenase, located in the mitochondrial outer membrane and also in lipid particles		
Q6CG32	YALI0B01298g	HFD3	YALI0B01298g1_1		Putative fatty aldehyde dehydrogenase	similar to uniprot|Q04458 Saccharomyces cerevisiae YMR110c HFD1 Putative fatty aldehyde dehydrogenase, located in the mitochondrial outer membrane and also in lipid particles		

Q6CGN3	YALI0A17875g	HFD4	YALI0A17875g1_1		Fatty aldehyde dehydrogenase	similar to uniprot|Q04458 Saccharomyces cerevisiae YMR110c HFD1 Putative fatty aldehyde dehydrogenase, located in the mitochondrial outer membrane and also in lipid particles		
Q6C704	YALI0E04807g	HMG1	YALI0E04807g1_1		HMG-CoA reductase	similar to uniprot|P12683 Saccharomyces cerevisiae YML075c HMG1 (ohnolog of YLR450W)		
Q6CGI2	YALI0A19096g	HTD2	YALI0A19096g1_1		Hydroxyacyl-thioester dehydratase	weakly similar to to uniprot|Q89E40 Bradyrhizobium japonicum Hypothetical cytosolic protein		
P41555	YALI0C16885g	ICL1	YALI0C16885g1_1	EC 4.1.3.1	Isocitrate lyase	uniprot|P41555 Yarrowia lipolytica YALI0C16885g Isocitrate lyase	Barth G and Scheuber T (1993) Mol Gen Genet 241(3-4):422-30; Forster et al (2007) Appl Microbiol Biotechnol. 77(4):861-9	Gene cloning and sequencing, Northern blots
Q6BZP5	YALI0F31999g	ICL2	YALI0F31999g1_1		Methylisocitrate lyase	similar to uniprot|Q12031 Saccharomyces cerevisiae YPR006c ICL2 2-methylisocitrate lyase of the mitochondrial matrix	Uchiyama H, Ando M, Toyonaka Y & Tabuchi T (1982) Eur J Biochem 125, 523–527	Subcellular localization of the methylcitric-acid-cycle enzymes
Q6C6Z1	YALI0E05137g	IDH1	YALI0E05137g1_1		isocitrate dehydrogenase (NAD+) subunit 1 mitochondrial	similar to uniprot|P28834 Saccharomyces cerevisiae YNL037c IDH1 isocitrate dehydrogenase (NAD+) subunit 1 mitochondrial		
Q6CA33	YALI0D06303g	IDH2	YALI0D06303g1_1		isocitrate dehydrogenase (NAD+) subunit 2 mitochondrial	highly similar to uniprot|P28241 Saccharomyces cerevisiae YOR136w IDH2 isocitrate dehydrogenase		
Q6C2Y4	YALI0F04095g	IDP1	YALI0F04095g1_1		isocitrate dehydrogenase	highly similar to uniprot|P41939 Saccharomyces cerevisiae YLR174w IDP2 isocitrate dehydrogenase		
Q6C016	YALI0F28567g	ISO1	YALI0F28567g1_1		enoyl-CoA hydratase/isomerase family protein	weakly similar to uniprot|Q9P3G4 Neurospora crassa conserved hypothetical protein		
Q6C9P1	YALI0D09493g	ISO2	YALI0D09493g1_1		enoyl-CoA hydratase/isomerase family protein	weakly similar to uniprot|Q9P3G4 Neurospora crassa Conserved hypothetical protein		
Q6CGD5	YALI0A20207g	ISO3	YALI0A20207g1_1		enoyl-CoA hydratase/isomerase family protein	some similarities with uniprot|Q9P3G4 Neurospora crassa Conserved hypothetical protein		
Q6C9P5	YALI0D09383g	ISO4	YALI0D09383g1_1		enoyl-CoA hydratase/isomerase family protein	similar to uniprot|Q9P3G4 Neurospora crassa Conserved hypothetical protein		
Q6C643	YALI0D00671g	ISO5	YALI0D00671g1_1		enoyl-CoA hydratase/isomerase family protein	similar to uniprot|Q9P3G4 Neurospora crassa Conserved hypothetical protein		
Q6C643	YALI0E12573g	IVD1	YALI0E12573g1_1		Isovaleryl-CoA dehydrogenase	similar to uniprot|B8NUG5 Aspergillus flavus AFLA_099270 isovaleryl-CoA dehydrogenase IvdA	Vorapreeda et al. (2012) Microbiology, 158, 217–228	
Q6C3M8	YALI0E33517g	KGD1	YALI0E33517g1_1		2-oxoglutarate dehydrogenase complex E1 component	highly similar to uniprot|P20967 Saccharomyces cerevisiae YIL125w KGD1 2-oxoglutarate dehydrogenase complex E1 component singleton	Holz M et al. (2011) Appl Microbiol Biotechnol 89:1519–1526	Gene overexpression
Q6C5L8	YALI0E16929g	KGD2	YALI0E16929g1_1		2-oxoglutarate dehydrogenase complex E2 component	similar to uniprot|P19262 Saccharomyces cerevisiae YDR148c KGD2 2-oxoglutarate dehydrogenase complex E2 component	Holz M et al. (2011) Appl Microbiol Biotechnol 89:1519–1526	Gene overexpression
Q6C812	YALI0D23683g	LAT1	YALI0D23683g1_1		Dihydrolipoamide acetyltransferase component (E2) of pyruvate dehydrogenase complex	similar to uniprot|P12695 Saccharomyces cerevisiae YNL071W LAT1 Dihydrolipoamide acetyltransferase		
Q6CEG0	YALI0B15950g	LDH1	YALI0B15950g1_1		member of the alpha/beta-hydrolase family	similar to CA2278|IPF10806 Candida albicans unknown function, hypothetical start		
Q99156	YALI0E10659g	LIP1	YALI0E10659g1_1		Carboxylesterase/lipase type B	uniprot|Q99156 Yarrowia lipolytica Lipase 1 precursor	Fickers P et al. (2011) Biotechnol. Adv. 29: 632-644	Electronic
Q6C225	YALI0F11429g	LIP10	YALI0F11429g1_1	EC 3.1.1.3	Triacylglycerol lipase precursor	weakly similar to uniprot|Q9P8F7 Yarrowia lipolytica Triacylglycerol lipase precursor	Fickers P et al. (2011) Biotechnol. Adv. 29: 632-644	Electronic
Q6C9R0	YALI0D09064g	LIP11	YALI0D09064g1_1	EC 3.1.1.3	Triacylglycerol lipase precursor	weakly similar to uniprot|Q9P8F7 Yarrowia lipolytica Triacylglycerol lipase precursor	Fickers P et al. (2011) Biotechnol. Adv. 29: 632-644	Electronic
I0CKV0	YALI0D15906g	LIP12	YALI0D15906g1_1	EC 3.1.1.3	Triacylglycerol lipase precursor	similar to uniprot|Q9P8F7 Yarrowia lipolytica Triacylglycerol lipase precursor	Fickers P et al. (2011) Biotechnol. Adv. 29: 632-644	Electronic
Q6C7J5	YALI0E00286g	LIP13	YALI0E00286g1_1	EC 3.1.1.3	Triacylglycerol lipase precursor	similar to uniprot|Q872L4 Yarrowia lipolytica LIP7 Lipase	Fickers P et al. (2011) Biotechnol. Adv. 29: 632-644	Electronic
Q6CEY5	YALI0B11858g	LIP14	YALI0B11858g1_1	EC 3.1.1.3	Triacylglycerol lipase precursor	similar to uniprot|Q872L4 Yarrowia lipolytica LIP7 Lipase	Fickers P et al. (2011) Biotechnol. Adv. 29: 632-644	Electronic
Q6C687	YALI0E11561g	LIP15	YALI0E11561g1_1	EC 3.1.1.3	Triacylglycerol lipase precursor	similar to uniprot|CAD70713 Yarrowia lipolytica lipase LIP4	Fickers P et al. (2011) Biotechnol. Adv. 29: 632-644	Electronic
Q6C8M4	YALI0D18480g	LIP16	YALI0D18480g1_1	EC 3.1.1.3	Triacylglycerol lipase precursor	similar to uniprot|Q872L4 Yarrowia lipolytica Lipase	Zhao H et al. (2011) Biotechnol. Lett. 33: 2445-2452;Fickers P et al (2011) Biotechnol. Adv. 29: 632-644	Cloning, expression in P. pastoris, specificity
Q6BZP1	YALI0F32131g	LIP17	YALI0F32131g1_1	EC 3.1.1.3	Triacylglycerol lipase precursor	similar to uniprot|CAD70715 Yarrowia lipolytica lipase	Fickers P et al. (2011) Biotechnol. Adv. 29: 632-644	Electronic
Q6CDY0	YALI0B20350g	LIP18	YALI0B20350g1_1	EC 3.1.1.3	Triacylglycerol lipase precursor	weakly similar to uniprot|Q9P8F7 Yarrowia lipolytica Triacylglycerol lipase precursor	Fickers P et al. (2011) Biotechnol. Adv. 29: 632-644	Electronic
Q6CHB8	YALI0A10439g	LIP19	YALI0A10439g1_1	EC 3.1.1.3	Triacylglycerol lipase precursor	similar to uniprot|Q9P8F7 Yarrowia lipolytica Triacylglycerol lipase precursor	Fickers P et al. (2011) Biotechnol. Adv. 29: 632-644	Electronic
Q9P8F7	YALI0A20350g	LIP2	YALI0A20350g1_1	EC 3.1.1.3	Triacylglycerol lipase precursor	uniprot|Q9P8F7 Yarrowia lipolytica Triacylglycerol lipase precursor	Pignede et al. (2000) J. Bacteriol. 182: 2802–2810	Gene cloning, sequencing, deletion
Q6C6V2	YALI0E05995g	LIP20	YALI0E05995g1_1		Carboxylesterase/lipase type B	weakly similar to uniprot|Q99156 Yarrowia lipolytica LIP1 Lipase 1 precursor	Fickers P et al. (2011) Biotechnol. Adv. 29: 632-644	Electronic
Q96VC9	YALI0B08030g	LIP3	YALI0B08030g1_1		Carboxylesterase/lipase type B	uniprot|Q96VC9 Yarrowia lipolytica Carboxylesterase/lipase type B	Fickers P et al. (2011) Biotechnol. Adv. 29: 632-644	Electronic
Q872L6	YALI0E08492g	LIP4	YALI0E08492g1_1	EC 3.1.1.3	Triacylglycerol lipase precursor	uniprot|Q872L6|emb|CAD70713 Yarrowia lipolytica YALI0E08492g LIP4 Lipase	Fickers P et al. (2011) Biotechnol. Adv. 29: 632-644	Electronic
CAD70714	YALI0E02640g	LIP5	YALI0E02640g1_1	EC 3.1.1.3	Triacylglycerol lipase precursor	uniprot|CAD70714 Yarrowia lipolytica Lipase	Fickers P et al. (2011) Biotechnol. Adv. 29: 632-644	Electronic
Q6CDI6	YALI0C00231g	LIP6	YALI0C00231g1_1		Carboxylesterase/lipase type B	similar to uniprot|Q99156 Yarrowia lipolytica Lipase 1 precursor	Fickers P et al. (2011) Biotechnol. Adv. 29: 632-644	Electronic
Q872L4	YALI0D19184g	LIP7	YALI0D19184g1_1	EC 3.1.1.3	Triacylglycerol lipase precursor	uniprot|Q872L4 Yarrowia lipolytica Lipase	Fickers P et al. (2005) Fungal Genet. Biol. 42: 264-274	Gene cloning, sequencing, deletion, substrate specificity
Q872L3	YALI0B09361g	LIP8	YALI0B09361g1_1	EC 3.1.1.3	Triacylglycerol lipase precursor	uniprot|Q872L3 Yarrowia lipolytica Lipase	Fickers P et al. (2005) Fungal Genet. Biol. 42: 264-274	Gene cloning, sequencing, deletion, substrate specificity
Q6C3I6	YALI0E34507g	LIP9	YALI0E34507g1_1	EC 3.1.1.3	Triacylglycerol lipase precursor	weakly similar to uniprot|Q9P8F7 Yarrowia lipolytica Triacylglycerol lipase precursor	Fickers P et al. (2011) Biotechnol. Adv. 29: 632-644	Electronic
Q6C8C6	YALI0D20768g	LPD1	YALI0D20768g1_1	EC 1.8.1.4	Dihydrolipoamide dehydrogenase	similar to uniprot|P09624 Saccharomyces cerevisiae YFL018c LPD1 (ohnolog of YPL017C) Dihydrolipoamide dehydrogenase	Morín M, Monteoliva L, Insenser M, Gil C, Domínguez A (2007) J Mass Spectrom 42(11):1453-62; Holz M et al. (2011) Appl Microbiol Biotechnol 89:1519–1526	Identified by 2D-gel analysis as associated to hyphae formation; gene overexpression
Q6C5M4	YALI0E16797g	LRO1	YALI0E16797g1_1		Phospholipid:diacylglycerol acyltransferase	similar to uniprot|P40345 Saccharomyces cerevisiae YNR008w LRO1 a lecithin cholesterol acyltransferase-like gene mediates diacylglycerol esterification singleton		
Q6C4S9	YALI0E24013g	LSC1	YALI0E24013g1_1		succinyl-CoA ligase alpha subunit	similar to uniprot|P53598 Saccharomyces cerevisiae YOR142w LSC1 succinate-CoA ligase alpha subunit		

Q6CA97	YALI0D04741g	LSC2	YALI0D04741g1_1		succinyl-CoA ligase beta subunit mitochondrial precursor	highly similar to uniprot|Q9P567 Neurospora crassa Probable succinyl-CoA ligase (GDP-forming) beta-chain mitochondrial precursor		
Q6C791	YALI0E02728g	LYS4	YALI0E02728g1_1		homoaconitase	highly similar to DEHA0F08019g Debaryomyces hansenii and uniprot|P49367 Saccharomyces cerevisiae YDR234w LYS4 homoaconitase	Holz M, Förster A, Mauersberger S, Barth G (2009) Appl Microbiol Biotechnol. 81: 1087-1096	Electronic
Q6C5F0	YALI0E18634g	MAE1	YALI0E18634g1_1	EC 1.1.1.40	Malic enzyme	similar to uniprot|P36013 Saccharomyces cerevisiae YKL029c MAE1 Mitochondrial malic enzyme	Zhang et al. (2013) Biotechnol lett. 35:2091-98; Blazeck et al. (2014) Nature Communications 5:3131 DOI: 10.1038/ncomms4131	Overexpression
Q6CEM0	YALI0B14619g	MCC1	YALI0B14619g1_1	EC 6.1.4.1	Methylcrotonoyl-CoA carboxylase, subunit alpha	similar to uniprot|E4UUI3 Arthroderma gypseum MGYG_03952 methylcrotonoyl-CoA carboxylase, subunit alpha		
Q6C642	YALI0E12595g	MCC2	YALI0E12595g1_1	EC 6.1.4.1	Methylcrotonoyl-CoA carboxylase, subunit beta	similar to uniprot|C5P540 CPC735_031660 Coccidioides posadasii methylcrotonyl CoA carboxylase beta chain, mitochondrial precursor	Vorapreeda et al. (2012) Microbiology, 158, 217–228	
Q6C5F2	YALI0E18590g	MCT1	YALI0E18590g1_1		Malonyl CoA-acyl carrier protein transacylase	similar to uniprot|Q9UUS5 Colletotrichum gloeosporioides Malonyl CoA-acyl carrier protein transacylase		
Q6C8V3	YALI0D16753g	MDH1	YALI0D16753g1_1	1.1.1.37	Mitochondrial malate dehydrogenase	highly similar to uniprot|Q8TG27 Talaromyces emersonii Malate dehydrogenase precursor, mitochondrial, start by similarity	Kabran P et al. (2012) DNA Res. 19(3):231-44. doi: 10.1093/dnares/dss007	Gene cloning, alternative splicing, protein localization
Q6C5X9	YALI0E14190g	MDH2lf	YALI0E14190g1_1	1.1.1.37	Cytoplasmic and peroxisomal malate dehydrogenase	similar to uniprot|P32419 Saccharomyces cerevisiae YDL078c MDH3 malate dehydrogenase, peroxisomal, start by similarity	Kabran P et al. (2012) DNA Res. 19(3):231-44. doi: 10.1093/dnares/dss007	Gene cloning, alternative splicing, protein localization
Q6C5X9	YALI0E14190g	MDH2sf	YALI0E14190g2_1	1.1.1.37	Cytoplasmic and peroxisomal malate dehydrogenase	similar to uniprot|P32419 Saccharomyces cerevisiae YDL078c MDH3 malate dehydrogenase, peroxisomal, start by similarity	Kabran P et al. (2012) DNA Res. 19(3):231-44. doi: 10.1093/dnares/dss007	Gene cloning, alternative splicing, protein localization
Q9P4D9	YALI0E15378g	MFE2	YALI0E15378g1_1		Multifunctional beta-oxidation enzyme, hydratase-dehydrogenase-epimerase, peroxisomal	uniprot|Q9P4D9 Yarrowia lipolytica Multifunctional beta-oxidation enzyme, hydratase-dehydrogenase-epimerase, peroxisomal, identified start	Smith JJ, Brown TW, Eitzen GA, Rachubinski RA (2000) J. Biol. Chem. 275:20168-20178	
Q6C0S5	YALI0F22121g	MGH1	YALI0F22121g1_1		Methylglutaconyl-CoA hydratase	similar to uniprot|A2QDJ5 Aspergillus niger methylglutaconyl-CoA hydratase		
Q6C5R9	YALI0E15708g	MLS1	YALI0E15708g1_1		Malate synthase	similar to uniprot|P21826 Saccharomyces cerevisiae YIR031c DAL7 Malate synthase		
Q6C8J4	YALI0D19140g	MLS2	YALI0D19140g1_1		Malate synthase	similar to uniprot|P30952 Saccharomyces cerevisiae YNL117w Malate synthase 1 glyoxysomal		
Q6CDN6	YALI0B22550g	MVA1	YALI0B22550g1_1		Hydroxymethylglutaryl-CoA lyase	similar to uniprot|A6QRM3 Ajellomyces capsulata mitochondrial hydroxymethylglutaryl-CoA lyase	Vorapreeda et al. (2012) Microbiology, 158, 217–228	
Q6CCJ7	YALI0C08811g	HBD1	YALI0C08811g1_1		3-hydroxyacyl-CoA dehydrogenase	similar to uniprot|Q9K6C9 Bacillus halodurans 3- hydroxybutyryl-CoA dehydrogenase		
Q6BZY3	YALI0F29975g	OAR1	YALI0F29975g1_1		3-oxoacyl-[acyl-carrier protein] reductase	weakly similar to uniprot|P38004 Chlamydia trachomatis 3- oxoacyl-[acyl-carrier protein] reductase		Electronic
Q6CCW6	YALI0C05951g	OLE1	YALI0C05951g1_1	EC 1.14.99.5	Delta-9 fatty acid desaturase	similar to uniprot|Q12618 Ajellomyces capsulata Delta-9 fatty acid desaturase (EC 1.14.99.5) (Acyl-CoA desaturase) (Stearoyl-CoA desaturase) (Delta(9)-desaturase)	Ferrante G et al. (1983) CAN. J. BIBCHEM. CELL BIOL. VOL. 61; Anamnart et al. (1997) Gene 184(2):299–306	
Q6C7L9	YALI0D27016g	PAH1	YALI0D27016g1_1		phosphatidate phosphatase	some similarities with uniprot|Q9UUJ6 Schizosaccharomyces pombe Nuclear elongation and deformation protein 1	Kobayashi S, Hirakawa K, Horiuchi H, Fukuda R and Ohta A (2013) Fungal Genet. Biol. 61, 100-110	Gene deletion
Q9C1T3	YALI0E11099g	PAT1	YALI0E11099g1_1	EC 2.3.1.9	Acetyl-CoA C-acetyltransferase (acetoacetyl-CoA thiolase)	uniprot|Q9C1T3 Yarrowia lipolytica Acetoacetyl-CoA thiolase (EC 2.3.1.9)	Yamagami S, Iida T, Nagata Y, Ohta A, Takagi M (2001) Biochem Biophys Res Commun 282(3):832-8	Cloning by complementation, mutagenesis, expression studies
Q6C0Y7	YALI0F20702g	PDA1	YALI0F20702g1_1		Pyruvate Dehydrogenase Alpha subunit	highly similar to uniprot|P16387 Saccharomyces cerevisiae YER178w PDA1 E1 alpha subunit of the pyruvate dehydrogenase		
Q6C9S7	YALI0D08690g	PDA2	YALI0D08690g1_1	EC 1.2.4.4	2-oxoisovalerate (lipoate) dehydrogenase E1 alpha chain	similar to uniprot|P50136 Mus musculus 2- oxoisovalerate dehydrogenase alpha subunit mitochondrial precursor		
Q6C4G4	YALI0E27005g	PDB1	YALI0E27005g1_1		Pyruvate Dehydrogenase Beta subunit	highly similar to uniprot|P32473 Saccharomyces cerevisiae YBR221c PDB1 E1 beta subunit of the pyruvate dehydrogenase		
Q6C2U7	YALI0F05038g	PDB2	YALI0F05038g1_1	EC 1.2.4.4	3-methyl-2-oxobutanoate dehydrogenase (Lipoamide) E1 beta chain	similar to uniprot|Q9P5K8 Neurospora crassa Probable 3- methyl-2-oxobutanoate dehydrogenase (Lipoamide)E1 beta chain	Vorapreeda et al. (2012) Microbiology, 158, 217–228	
Q6C9L5	YALI0D10131g	PDC1	YALI0D10131g1_1		Pyruvate decarboxylase	similar to uniprot|P06169 Saccharomyces cerevisiae YLR044c PDC1 pyruvate decarboxylase isozyme 1		Electronic
Q12428	YALI0F02497g	PDH1	YALI0F02497g1_1		2-methylcitrate dehydratase	similar to uniprot|A4D9B3 Aspergillus fumigatus 2-methylcitrate dehydratase (PrpD)	Uchiyama H, Ando M, Toyonaka Y & Tabuchi T (1982) Eur J Biochem 125, 523–527; Papanikolaou et al. (2013) J Biotechnol 4(168):303-314	Gene deletion
Q6CF67	YALI0B09845g	PDX1	YALI0B09845g1_1		pyruvate dehydrogenase complex protein X	similar to uniprot|P16451 Saccharomyces cerevisiae YGR193c PDX1 pyruvate dehydrogenase complex protein X, start by similarity		
Q6C9A9	YALI0D12628g	POR1	YALI0D12628g1_1		Subunit of the SWI/SNF chromatin remodeling complex involved in transcriptional regulation	uniprot|Q6C9A9 Yarrowia lipolytica, transcriptional activator of fatty acid utilization, homologue of A. nidulans FarA	Poopanitpan N et al. (2010) BBRC 402:731–735	
Q05493	YALI0E18568g	POT1	YALI0E18568g1_1		3-ketoacyl-CoA thiolase with broad chain length specificity	uniprot|Q05493 Yarrowia lipolytica 3-ketoacyl-CoA thiolase peroxisomal precursor		
O74934	YALI0E32835g	POX1	YALI0E32835g1_1	EC 1.3.3.6	Fatty-acyl coenzyme A oxidase	uniprot|O74934 Yarrowia lipolytica POX1 Acyl-CoA oxidase 1 (EC 1.3.3.6), peroxisomal, specific toward dicarboxilic acid		
O74935	YALI0F10857g	POX2	YALI0F10857g1_1	EC 1.3.3.6	Fatty-acyl coenzyme A oxidase	uniprot|O74935 Yarrowia lipolytica POX2 Acyl-CoA oxidase 2 (EC 1.3.3.6)		
O74936	YALI0D24750g	POX3	YALI0D24750g1_1	EC 1.3.3.6	Fatty-acyl coenzyme A oxidase	uniprot|O74936 Yarrowia lipolytica POX3 Acyl-CoA oxidase 3 (EC 1.3.3.6), peroxisomal, specific toward short chain fatty acid		
O74937	YALI0E27654g	POX4	YALI0E27654g1_1	EC 1.3.3.6	Fatty-acyl coenzyme A oxidase	uniprot|O74937 Yarrowia lipolytica POX4 Acyl-CoA oxidase 4 (EC 1.3.3.6), peroxisomal		
O74938	YALI0C23859g	POX5	YALI0C23859g1_1	EC 1.3.3.6	Fatty-acyl coenzyme A oxidase	uniprot|O74938 Yarrowia lipolytica POX5 Acyl-CoA oxidase 5 (EC 1.3.3.6), peroxisomal		
Q6C6T0	YALI0E06567g	POX6	YALI0E06567g1_1	EC 1.3.3.6	Fatty-acyl coenzyme A oxidase	similar to uniprot|O74935 Yarrowia lipolytica YlPOX2 - YALI0F10857g Acyl-CoA oxidase 2		
Q6CHP7	YALI0A06655g	PXA1	YALI0A06655g1_1		Subunit of a heterodimeric peroxisomal ATP-binding cassette transporter complex (Pxa1p-Pxa2p)	similar to uniprot|P41909 Saccharomyces cerevisiae YPL147w PXA1 Subunit of a heterodimeric peroxisomal ATP-binding cassette transporter complex (Pxa1p-Pxa2p), required for import of long-chain fatty acids into peroxisomes		

Q6CAB6	YALI0D04246g	PXA2	YALI0D04246g1_1		Subunit of a heterodimeric peroxisomal ATP-binding cassette transporter complex (Pxa1p-Pxa2p)	similar to uniprot|P34230 Saccharomyces cerevisiae YKL188c PXA2 Subunit of a heterodimeric peroxisomal ATP-binding cassette transporter complex (Pxa1p-Pxa2p), required for import of long-chain fatty acids into peroxisomes		
Q6C9L5	YALI0C24101g	PYC1	YALI0C24101g1_1		Pyruvate carboxylase	similar to uniprot|P06169 Saccharomyces cerevisiae YLR044c PDC1 pyruvate decarboxylase isozyme 1	Flores CL and Gancedo C (2005) Euk. Cell 4: 356-364	Gene deletion
P30614	YALI0F09185g	PYK1	YALI0F09185g1_1		Pyruvate kinase	uniprot|P30614 Yarrowia lipolytica YALI0F09185g Pyruvate kinase	Strick CA et al. (1992) Gene 118: 65-72	Cloning, amplification, intron mapping
Q6CDI7	YALI0C00209g	SCT1	YALI0C00209g1_1		glycerol-3-phosphate O-acyltransferase	similar to uniprot|P32784 Saccharomyces cerevisiae YBL011w SCT1 suppresses a choline-transport mutant		
Q6C9G6	YALI0D11374g	SDH1	YALI0D11374g1_1	EC 1.3.5.1	succinate dehydrogenase	similar to uniprot|P47052 Saccharomyces cerevisiae Probable succinate dehydrogenase ubiquinone flavoprotein subunit 2 mitochondrial precursor (EC 1.3.5.1)(Flavoprotein subunit of complex II) and uniprot|Q00711 Saccharomyces cerevisiae YKL148c SDH1 succinate dehydrogenase flavoprotein precursor P2.302.f2.1		
Q6C823	YALI0D23397g	SDH2	YALI0D23397g1_1	EC 1.3.5.1	succinate dehydrogenase	highly similar to uniprot|P21801 Saccharomyces cerevisiae YLL041C Succinate dehydrogenase [ubiquinone] iron-sulfur protein mitochondrial precursor (EC 1.3.5.1) (IP)		
Q6C450	YALI0E29667g	SDH3	YALI0E29667g1_1	EC 1.3.5.1	succinate dehydrogenase	weakly similar to uniprot|P33421 Saccharomyces cerevisiae YKL141w SDH3 cytochrome b560 subunit of respiratory complex II		
Q6CGZ0	YALI0A14784g	SDH4	YALI0A14784g1_1	EC 1.3.5.1	succinate dehydrogenase	some similarities with uniprot|P37298 Saccharomyces cerevisiae YDR178w SDH4 succinate dehydrogenase membrane anchor subunit for SDH2P		
Q6C098	YALI0F26587g	SKT1	YALI0F26587g1_1		succinyl-CoA:3-ketoacid-coenzyme A transferase	similar to CA0994|IPF10077 Candida albicans		
Q6C5D5	YALI0E18964g	SLC1	YALI0E18964g1_1		fatty acyltransferase	similar to uniprot|P33333 Saccharomyces cerevisiae YDL052c SLC1 fatty acyltransferase		
Q6CAL6	YALI0D01694g	SPS191	YALI0D01694g1_1	1.3.1.34	Peroxisomal 2,4-dienoyl-CoA reductase	similar to uniprot|P32573 Saccharomyces cerevisiae YNL202w SPS19 Peroxisomal 2,4-dienoyl-CoA reductase, auxiliary enzyme of fatty acid beta-oxidation		
Q6C5X5	YALI0E14322g	SPS192	YALI0E14322g1_1	1.3.1.34	Peroxisomal 2,4-dienoyl-CoA reductase	similar to uniprot|P32573 Saccharomyces cerevisiae YNL202w SPS19 Peroxisomal 2,4-dienoyl-CoA reductase, auxiliary enzyme of fatty acid beta-oxidation		
Q6CD80	YALI0C03003g	SPS193	YALI0C03003g1_1	1.3.1.34	Peroxisomal 2,4-dienoyl-CoA reductase	similar to uniprot|P32573 Saccharomyces cerevisiae YNL202w SPS19 Peroxisomal 2,4-dienoyl-CoA reductase, auxiliary enzyme of fatty acid beta-oxidation		
Q6C1N6	YALI0F14729g	TES1	YALI0F14729g1_1		peroxisomal Acyl-CoA thioesterase	similar to uniprot|Q9C3Z1 Cochliobolus heterostrophus peroxisomal Acyl-CoA thioesterase, start by similarity		
Q6C3U6	YALI0E32035g	TGL1	YALI0E32035g1_1	3.1.1.3	triacylglycerol lipase	weakly similar to uniprot|P34163 Saccharomyces cerevisiae YKL140w TGL1 triacylglycerol lipase P3.90.f3.1		
Q6C3X3	YALI0E31515g	TGL2	YALI0E31515g1_1	3.1.1.3	triacylglycerol lipase	similar to CA5070|IPF3594 Candida albicans IPF3594 triglyceride lipase (by homology)		
Q6C8R8	YALI0D17534g	TGL3	YALI0D17534g1_1	3.1.1.3	triacylglycerol lipase	similar to DEHA0G23430g Debaryomyces hansenii IPF 4879.1		
Q6C8W4	YALI0D16379g	TGL32	YALI0D16379g1_1	3.1.1.3	triacylglycerol lipase	some similarities with uniprot|O14115 Schizosaccharomyces pombe Hypothetical protein		
Q6C282	YALI0F10010g	TGL4	YALI0F10010g1_1	3.1.1.3	triacylglycerol lipase	weakly similar to uniprot|P36165 Saccharomyces cerevisiae YKR089c and DEHA0E06127g Debaryomyces hansenii		
Q6L8K6	YALI0C02387g	YAS1	YALI0C02387g1_1		basic helix-loop-helix transcription factor essential for cytochrome p450 induction in response to alkanes	uniprot|Q6L8K6 Yarrowia lipolytica YALI0C02387g YAS1 basic helix-loop-helix transcription factor essential for cytochrome p450 induction in response to alkanes	Yamagami S et al. (2004) JBC 279:22183-22189	Cloning, mutation, localization, promoter binding assays
Q6C3S8	YALI0E32417g	YAS2	YALI0E32417g1_1		basic helix-loop-helix transcription factor essential for cytochrome p450 induction in response to alkanes	uniprot|Q6C3S8 Yarrowia lipolytica YALI0C02387g YAS2 basic helix-loop-helix transcription factor essential for cytochrome p450 induction in response to alkanes	Endoh-Yamagami S (2007) Eukaryot Cell. 6(4):734-43	Gene deletion, binding to Yas1p, effect on ALK transcription
B9X0I3	YALI0C14784g	YAS3lf	YALI0C14784g1_1		long form of transcription repressor of ALK genes, expressed on glucose, Opi1 homologue	uniprot|B9X0I3 Yarrowia lipolytica long form of transcription repressor YlYAS3p of ALK genes, expressed on glucose, Opi1 homologue	Hirakawa K, Kobayashi S, Inoue T, Endoh-Yamagami S, Fukuda R, Ohta A. J Biol Chem. 2009 284(11):7126-37	RACE, gene deletion, binding to Yas2p, subcellular localization, effect on ALK genes expression
B9X0I3	YALI0C14784g	YAS3sf	YALI0C14784g2_1		short form of transcription repressor of ALK genes, induced on alkanes, Opi1 homologue	uniprot|B9X0I3 Yarrowia lipolytica short form of transcription repressor YlYAS3p of ALK genes, induced on alkanes, Opi1 homologue	Hirakawa K, Kobayashi S, Inoue T, Endoh-Yamagami S, Fukuda R, Ohta A. J Biol Chem. 2009 284(11):7126-37	RACE, gene deletion, binding to Yas2p, subcellular localization, effect on ALK genes expression
Q6C0W4	YALI0F21197g	YAT1	YALI0F21197g1_1		Outer mitochondrial carnitine acetyltransferase	similar to uniprot|Q8TGF1 Aspergillus fumigatus Carnitine acetyl transferase		
Q6CF30	YALI0B10736g	YHM2	YALI0B10736g1_1		Citrate and oxoglutarate carrier protein	highly similar to uniprot|Q04013 Saccharomyces cerevisiae YMR241w YHM2 yeast suppressor gene of HM (mitochondrial histone) mutant (ABF2) singleton		
Q6C5J8	YALI0E17457g	YMC1	YALI0E17457g1_1		Mitochondrial organic acid transmembrane transporter	similar to uniprot|P32331 Saccharomyces cerevisiae YPR058w YMC1 (ohnolog of YBR104W) Mitochondrial protein, putative inner membrane transporter with a role in oleate metabolism and glutamate biosynthesis		
Q6C4B3	YALI0E28270g	YMC11	YALI0E28270g1_1		Mitochondrial organic acid transmembrane transporter	similar to uniprot|P32331 Saccharomyces cerevisiae YPR058w YMC1 mitochondrial carrier protein (MCF)		
Q6CGA0	YALI0A20988g	YMC2	YALI0A20988g1_1		Mitochondrial organic acid transmembrane transporter	similar to CA5146|CaYMC2 Candida albicans CaYMC2 Carnitine/acylcarnitine translocase		

Dynamique des génomes et évolution du métabolisme lipidique chez les levures du clade *Yarrowia*

Yarrowia lipolytica appartient à un groupe de levures qui a divergé très tôt dans l'arbre des hémiascomycètes. Cette levure est capable d'utiliser, comme seule source de carbone, des substrats hydrophobes très variés et est capable de synthétiser de nouveaux acides gras libres à partir de composés non hydrophobes. Ces caractéristiques font de *Y. lipolytica* un modèle de levure hémiascomycète pour l'étude du métabolisme des lipides. Sa capacité oléagineuse est facilitée par l'expansion de familles de protéines ayant eu lieu au cours de l'évolution après la divergence du clade *Yarrowia*. En effet, parmi les 204 gènes impliqués dans le métabolisme des lipides dans *Y. lipolytica*, plus de 67% sont regroupés dans 30 familles multigéniques avec un maximum de 16 membres pour la famille de la lipase.
Le récent séquençage de génomes 5 dans le clade *Yarrowia* a permis de comparer leur contenu génétique. L'étude des familles de gènes du métabolisme des lipides a permis de mettre en évidence les mécanismes impliqués dans l'évolution de ce clade. Toutes les fonctions nécessaires pour le métabolisme des lipides sont présentes dans toutes les espèces étudiées, avec au moins un gène par famille, même pour les espèces qui ont de 500 à 1000 moins de gènes que *Y. lipolytica*. Ces espèces se sont d'ailleurs révélées être toutes oléagineuses, grâce à des études physiologiques.
Les familles de gènes observées proviennent de multiples duplications de gènes dont chaque exemplaire évolue indépendamment par différents processus de sub- ou neofonctionalisations, fixations, pseudogénisations ou pertes. Les contractions et expansions de familles de protéines, l'expression des gènes, la synténie et la localisation relative des gènes dans le génome ont été étudiés. Plus particulièrement, la pression de sélection agissant sur ces familles de gènes a été comparée à celles agissant sur le reste du génome.
La combinaison de ces approches, physiologie, génomique et transcriptomique, a permis d'améliorer la compréhension du métabolisme lipidique, de son évolution et de la régulation des gènes dans un clade des levures oléagineuses.

Genome dynamics and evolution of the lipid metabolism in yeasts of the *Yarrowia* clade

Yarrowia lipolytica belongs to a group of yeasts that have diverged very early from most other hemiascomycetous yeasts. This yeast is able to use various hydrophobic substrates as unique carbon source and to synthesize new free fatty acid from non hydrophobic compounds. These characteristics make *Y. lipolytica* a known oleaginous model for the lipid metabolism survey of yeasts. Its oleaginous capacity is facilitated by protein family expansions that occurred across evolution after the divergence of the *Yarrowia* clade. Indeed, among 190 genes involved in the lipid metabolism in *Y. lipolytica*, more than 67% are grouped into 30 multigenic families with up to 16 members in the lipase family.
The recent sequencing of 5 genomes within the *Yarrowia* clade enabled to compare their gene content. The study of gene families of the lipid metabolism allowed to highlight the evolutionary mechanisms involved in this clade. All the functions necessary to lipid metabolism are present in all species studied with at least one gene per family even for species that have 500 to 1,000 fewer genes than *Y. lipolytica*. These species are also found to be all oleaginous through physiological studies.
The observed gene families derive from multiple gene duplications of which each copy evolves independently by different processes of sub- or neofunctionalisation, fixation, pseudogenization or loss. Contractions and expansions of protein families, gene expression, synteny and relative localisation of genes in the genome were investigated. More particularly, the selection pressure acting on these gene families was compared with those acting on the rest of the genome.
The combination of these approaches, physiology, genomics and transcriptomics, has improved the comprehension of the lipid metabolism, its evolution and gene regulation within a clade of oleaginous yeasts.