Statistical modelling for differential gene expression studies: variance-covariance models, sequential and meta-analysis.

Guillemette Marot

N° /2009/AGPT/0039/

THÈSE

pour obtenir le grade de

Docteur

de

l'Institut des Sciences et Industries du Vivant et de l'Environnement (Agro Paris Tech)

Spécialité : Mathématiques appliquées

présentée et soutenue publiquement par

Guillemette MAROT

le 9 septembre 2009

MODELISATION STATISTIQUE POUR LA RECHERCHE DE GENES DIFFERENTIELLEMENT EXPRIMES:

MODELES DE VARIANCE-COVARIANCE, ANALYSE SEQUENTIELLE ET META-ANALYSE

Directeurs de thèse : ***Florence JAFFREZIC/ Jean-Louis FOULLEY/ Claus-Dieter MAYER***

Travail réalisé :
INRA, UMR1313 Génétique Animale et Biologie Intégrative, F-78350 Jouy-en-Josas, France
BioSS, Rowett Institute, UK-AB21 9SB Aberdeen, Scotland

Devant le jury :

M. Gilles CELEUX, Directeur de recherches, **INRIA**……..**Président**
M. Philippe BESSE, Professeur, **Université de Toulouse** …….. **Rapporteur**
M. Jean-Louis FOULLEY, Directeur de recherches, **INRA** …… **Examinateur**
M. Thomas HEAMS, Maître de Conférences, **Agro Paris Tech** …….. **Examinateur**
Mme Florence JAFFREZIC, Chargée de recherches, **INRA** …… **Examinateur**
M. Stéphane ROBIN, Directeur de recherches, **INRA** …… **Examinateur**
M. Korbinian STRIMMER, Professeur, **Université de Leipzig** …….. **Rapporteur**

INRA, Génétique Animale et Biologie Intégrative
Batiment 211 - Domaine de Vilvert
78350 Jouy-en-Josas
France

Biomathematics and Statistics Scotland
Rowett Institute
Bucksburn
Aberdeen AB21 9SB
UK Scotland

Modélisation statistique pour la recherche de gènes différentiellement exprimés: modèles de variance-covariance, analyse séquentielle et méta-analyse

Les puces à ADN permettent d'étudier simultanément l'expression de plusieurs milliers de gènes à partir de peu d'individus biologiques. Trois approches sont considérées dans cette thèse pour résoudre les problèmes de sensibilité dans la recherche de gènes différentiellement exprimés: la modélisation des variances-covariances, l'analyse séquentielle et la méta-analyse. La première et la troisième partie reposent principalement sur des approches dites de 'shrinkage' qui estiment les valeurs de chaque gène à partir de l'information provenant de l'ensemble des gènes. En diminuant le nombre de paramètres à estimer, elles permettent d'augmenter la sensibilité. La modélisation des variances se révèle particulièrement utile dans le cas d'expériences avec de petits échantillons. La modélisation des covariances est quant à elle particulièrement pertinente pour les études de suivi longitudinal où les mesures sont répétées sur les mêmes individus au cours du temps. Côté analyse séquentielle, la sensibilité est étudiée en tant que règle d'arrêt. On cherche alors à arrêter une expérience en cours dès que ce critère dépasse un certain seuil, afin d'en diminuer les coûts. La méta-analyse est ensuite étudiée dans un contexte beaucoup plus général que celui de l'analyse séquentielle où on combinait les analyses intermédiaires. Elle permet de gagner de la sensibilité en regroupant des résultats d'études individuelles qui ne sont pas comparables directement mais qui répondent à une même question biologique. La méta-analyse est abordée à la fois sous l'angle fréquentiste (combinaison de grandeurs des effets ou combinaison de p-values) et sous l'angle bayésien.

Mots clefs: puces à ADN, analyse différentielle, modélisation des variances-covariances, dépendance au cours du temps, analyse séquentielle, méta-analyse.

Statistical modelling for differential gene expression studies: variance-covariance models, sequential and meta-analysis.

Microarrays enable to simultaneously study gene expression levels from several thousands of genes with very few samples. Three approaches are considered in this PhD work in order to overcome sensitivity problems in differential gene expression studies: variance-covariance modelling, sequential and meta-analysis. The first and the third parts mainly rely on shrinkage approaches, which consist in estimating each individual gene value by taking into account information from all genes of the experiment. By decreasing the total number of parameters to estimate, this increases sensitivity, that is to say the proportion of true positives among the truly differentially expressed genes. While variance modelling is always useful with small sample size designs, covariance modelling is especially important in time course studies where measures are repeated on the same individuals. Concerning sequential analysis, sensitivity is studied as a stopping rule. The aim is to stop the experiment before the scheduled end as soon as this criterion is higher than a given threshold, which enables to decrease costs. Meta-analysis is then studied in a wider context than sequential analysis where intermediate analyses were combined. It increases sensitivity by gathering results from individual studies, for which a direct comparison would be impossible, but answering the same biological question. Meta-analysis is studied both from the frequentist (effect size and p-value combinations) and the bayesian points of view.

Keywords: microarrays, differential analysis, variance-covariance modelling, time course studies, sequential analysis, meta-analysis.

Remerciements/Acknowledgements

First of all, I would like to thank people who have accepted to judge my PhD work, especially Korbinian Strimmer and Philippe Besse, who will review my report, but also Gilles Celeux and Stéphane Robin, who have agreed to be examiners at my PhD defense.

Je tiens ensuite à remercier Didier Boichard et Philippe Chemineau, Chefs des Départements GA (Génétique Animale) et PHASE (Physiologie et Systèmes d'Elevage) qui ont permis le financement de ma thèse.

Je remercie Jean-Pierre Bidanel et toute l'équipe de l'ex-Station de Génétique Quantitative et Appliquée, maintenant intégrée dans la très grande unité Génétique Animale et Biologie Intégrative (GABI), pour son accueil chaleureux. Merci aussi aux membres de ma nouvelle équipe Populations, Statistique et Génomique.

Je remercie plus particulièrement Florence Jaffrézic et Jean-Louis Foulley qui ont initié et dirigé ma thèse pendant ces trois années. J'ai beaucoup apprécié leur grande complémentarité. Je remercie Florence pour ses conseils, son attention particulière à me faire rencontrer les bonnes personnes au moment où il le fallait. Grâce à elle, j'ai tiré beaucoup d'avantages des congrès auxquels j'ai participé. Merci d'avoir su me motiver quand les résultats n'étaient pas aussi concluants que je l'aurais voulu et merci de m'avoir guidée vers mon futur post-doctorat à Lyon avec Franck Picard.

Je remercie aussi très sincèrement Jean-Louis Foulley pour son aide précieuse dans des situations délicates au niveau théorique, ses bonnes idées scientifiques. Je le remercie de m'avoir fait découvrir la statistique bayésienne et d'avoir consacré une partie de son temps à essayer divers modèles pour comprendre pourquoi mes résultats n'étaient pas concluants pendant ma dernière ligne droite de thèse. J'ai beaucoup d'autres personnes à remercier dans GABI mais avant de les nommer, je voudrais remercier mon autre co-directeur de thèse: Claus-Dieter Mayer.

I would like to say a special thank you to my co-supervisor Claus-Dieter Mayer. It has been a great pleasure to work with him, both as an expert in his field and a very nice supervisor. Although my longest stay in Scotland lasted only six months, he co-supervised me during more than two years during my PhD. After my stay in Aberdeen at the end of the first year of my PhD, we stayed in touch with regular phone conversations regarding my research, and I learned a great deal from him about microarrays and statistics. Although he was very busy, he always found time to help me, both on methodological and practical questions. I was impressed by the quality of his remarks, which

greatly improved my papers and my PhD project. He also provided a warm welcome for me in Scotland and thanks to him and his wife Janine, I spent a lovely time there and was able to improve my English. Of course, I am still a bit disappointed that Claus cannot be an examiner at my PhD's defence but I completely understand the position of the doctoral school that all three supervisors cannot be present in my judging comittee. However, I will not forget the important role he played in my PhD work.

I would like to jointly thank these three supervisors for their flexibility in accepting compromises when I was asked double or triple corrections, especially when changes were not necessarily unanimous. They were all able to differentiate which fields were their expertise or not, which explains the success of my 3-person cosupervision. Many thanks to them for that!

I would also like to thank the people at Biomathematics and Statistics Scotland who contributed to making my stay in Scotland enjoyable: Chris Glasbey who accepted the INRA-BioSS collaboration and the financing of life expenses during my six month stay, Graham Horgan the leader of BioSS team in Aberdeen and Grietje Holtrop who both made sure with Claus that I had everything I needed. From the Rowett Institute where BioSS team was housed, I especially thank Tony Travis, who let me use his cluster.

Côté français, j'aimerais remercier Luc Jouneau, bioinformaticien avec qui cela a été un véritable plaisir de travailler sur sa chaîne d'analyse statistique pour les puces à ADN. Merci aussi à tous les biologistes qui m'ont fourni des jeux de données de bonne qualité et des questions très intéressantes: Isabelle Hue, Séverine Degrelle, Benoit Guyonnet, Jean-Luc Gatti, Damien Valour, Olivier Sandra.

Je remercie particulièrement Isabelle Hue car elle a aussi aidé Florence et Jean-Louis dans l'initiation de ce projet de thèse et m'a suivi très attentivement lors de mes comités de thèse. J'en profite pour remercier les autres membres de mon comité de thèse: Jean-Jacques Daudin que je remercie aussi de m'avoir recommandée auprès de Franck Picard, Thomas Heams qui a bien rempli son rôle de tuteur de l'école doctorale ABIES (en suppléant Etienne Verrier), Isabel Brito et Olivier Martin pour toutes leurs remarques constructives.

Merci à l'école doctorale pour toutes les formations de qualité auxquelles j'ai pu participer.

Je remercie aussi les professeurs et maîtres de conférence que j'ai rencontrés à l'IUT Paris V. Merci à Adeline Samson de m'avoir proposé de donner des TD là-bas, un grand merci à Florence Muri-Majoube et Fanny Villers

avec qui j'ai été particulièrement heureuse de travailler pour les cours de statistique descriptive et séries chronologiques, merci à Guillaume Bordry pour ses emplois du temps arrangeants et sa sympathie. Merci aussi à Allou Same et Marie-Luce Taupin que j'ai vus un peu moins souvent mais dont j'ai aussi eu l'opportunité de partager les matières et les copies. Et merci aussi à mes élèves!

Merci à tous les membres du groupe de travail Statomique pour les échanges très intéressants.

Par ailleurs, je tiens à ne pas oublier François Guillaume, Hervé Lagant et Fabien Dequine pour leur grande disponibilité, leurs astuces et leurs dépannages informatiques, Wendy pour la relecture de l'anglais de mes articles. Merci aussi à Manuéla, Sylvie, Christelle, Pascale, Serge et François pour leur travail administratif et de reprographie.

Il y aurait beaucoup d'autres personnes que j'aimerais remercier pour leur soutien mais je pense qu'il faut quand même un peu écourter. Cependant, je tiens à remercier mes actuelles voisines de bureau Andrea Rau et Sandrine Schwob qui ont été particulièrement attentionnées pendant ma dernière ligne droite de thèse. Je suis très redevable à Andrea pour ses conseils surtout anglophones mais aussi parfois statistiques. Merci aussi à tous les anciens et nouveaux thésards que je n'ai pas encore remercié ainsi qu'à tous les stagiaires, particulièrement ceux que j'ai cotoyé le plus longtemps: Lauriane Canario, Florence Ytournel, Hélène Leclerc, Alban Bouquet, Olympe Chazara, Sophie Allais, Btissam Salmi, Clotilde Patry.

Of course, I cannot forget all the students or post-docs that I met in Scotland, especially those who still give me regular news: Laura, Myrte, Cindy, Elodie and Pierre-Emmanuel.

Merci aussi à tous mes amis. La liste serait trop longue pour énumérer chacun de ceux qui m'ont soutenue mais j'ai une pensée particulière pour mes coloc' qui m'ont supportée ces dernières années ainsi que ceux de mes amis qui ont été très présents aux moments moins faciles de cette thèse. Enfin, je ne peux finir sans remercier mes parents et ma soeur jumelle qui ont été d'un soutien exemplaire tout du long.

Résumé substantiel

Les études transcriptomiques basées sur des expériences de puces à ADN sont devenues un outil standard dans les sciences de la vie au cours de ces dix dernières années. Cependant, le coût de ces expériences reste élevé, ce qui se traduit souvent par un manque d'échantillons disponibles. Ma thèse étudie le problème de dimension élevée (très grand nombre de gènes/variables) associée à un très faible nombre d'échantillons impliqués dans la recherche de gènes différentiellement exprimés. Trois approches principales sont considérées: la modélisation des variances-covariances, l'analyse séquentielle et la méta-analyse.

Shrinkage

Les modélisations des variances-covariances et la méta-analyse proposées ici sont basées sur des approches dites de shrinkage. Ce mot clef est utilisé à chaque fois qu'un estimateur est un compromis entre deux estimateurs. De manière générale, l'estimateur de shrinkage $\tilde{\boldsymbol{\theta}}_{\boldsymbol{g}}$ peut s'écrire comme une fonction d'un estimateur gène à gène $\widehat{\boldsymbol{\theta}_{\boldsymbol{g}}}$ et d'un estimateur commun de la population globale $\widehat{\boldsymbol{\theta}_{\boldsymbol{c}}}$:

$$\tilde{\boldsymbol{\theta}}_{\boldsymbol{g}} = \widehat{\boldsymbol{\theta}_{\boldsymbol{c}}} + b(\widehat{\boldsymbol{\theta}_{\boldsymbol{g}}} - \widehat{\boldsymbol{\theta}_{\boldsymbol{c}}}) \tag{1}$$

où b est le facteur de shrinkage. Quand $b = 1$, $\tilde{\boldsymbol{\theta}}_{\boldsymbol{g}} = \widehat{\boldsymbol{\theta}_{\boldsymbol{g}}}$ (estimateur empirique gène à gène). Quand $b = 0$, $\tilde{\boldsymbol{\theta}}_{\boldsymbol{g}} = \widehat{\boldsymbol{\theta}_{\boldsymbol{c}}}$ (estimateur commun). Les approches de shrinkage diminuent considérablement le nombre de paramètres à estimer tout en gardant une certaine flexibilité avec une valeur par gène.

Modélisation des variances-covariances

La modélisation de la variance joue un rôle important dans les études de gènes différentiellement exprimés. Quand très peu d'échantillons sont considérés et que l'analyse est réalisée gène à gène, les tests statistiques manquent de puissance; dans ce contexte, cela veut dire que très peu de gènes différentiellement exprimés peuvent être détectés. Une alternative est de supposer une variance commune à tous les gènes. Cependant, cela conduit souvent à une augmentation du nombre de faux positifs. Au cours de mon stage de fin d'étude de l'Ecole Nationale de la Statistique et de l'Analyse de l'Information (ENSAI), j'ai programmé une nouvelle approche de shrinkage basée sur le modèle structural mixte pour les variances. Dans ce cas, θ de l'équation (2) correspond au log des variances et s'écrit comme un modèle mixte avec un effet condition fixe et un effet gène aléatoire. Le facteur de

shrinkage est estimé via une approche bayésienne empirique. Au début de ma thèse, j'ai étendu le modèle de variance à la modélisation de matrices de covariance, ce qui est particulièrement intéressant pour les études au cours du temps quand les mesures sont répétées sur les mêmes individus. Avec F. Jaffrézic et J.-L. Foulley, nous avons proposé une approche de shrinkage basée sur un modèle structural mixte via une décomposition en valeurs propres ou une décomposition de Cholesky des matrices de variance-covariance. Les estimateurs de shrinkage ont été calculés sur trois niveaux i) les valeurs propres ii) les variances d'innovation iii) à la fois les variances et les paramètres de corrélation d'une matrice de corrélation empirique gène à gène. Nous avons trouvé que les méthodes proposées se comportaient bien par rapport aux approches Bayésiennes empiriques déjà existantes et étaient meilleures, dans la plupart des cas, que les méthodes gène à gène ou les approches supposant une covariance commune.

Analyse séquentielle

L'analyse séquentielle considère le problème de taille d'échantillonage d'un autre point de vue. L'idée pour réduire les coûts des expériences est d'être capable d'arrêter une expérience une fois que suffisamment de résultats ont été obtenus. Les approches séquentielles ont une longue tradition dans les essais cliniques pour réduire la taille d'échantillonnage tout en gardant une puissance statistique raisonnable. Ces méthodes sont caractérisées par des analyses intermédiaires à des étapes pré-définies et une règle d'arrêt qui détermine à chaque étape si on doit continuer l'échantillonnage ou non. Au cours d'un séjour en Ecosse pendant ma thèse, j'ai proposé avec C.-D. Mayer (BioSS) une approche séquentielle pour les puces à ADN. Une caractéristique intéressante d'une telle approche est que, contrairement au cas univarié, le grand nombre de variables (gènes) testées simultanément empêche l'introduction d'un biais considérable des p-values finales. Ainsi, les résultats des différentes étapes peuvent être combinés par des méthodes de méta-analyse et les taux d'erreurs contrôlés en appliquant les procédures classiques (par exemple correction de Benjamini Hochberg) aux p-values résultant de la combinaison des différentes étapes. Nous avons proposé des règles d'arrêt basées soit sur l'estimation du nombre de vrais positifs soit sur l'estimation de la sensibilité, c'est-à-dire la proportion de vrais positifs parmi les gènes réellement différentiellement exprimés. Nous avons comparé plusieurs modèles de mélange pour estimer la sensibilité et montré qu'il était difficile d'estimer ce critère. Grâce à des simulations, nous avons aussi trouvé que les approches séquentielles étaient capables de réduire les tailles d'échantillonnage et par conséquent les coûts dans les expériences de puces à ADN.

Méta-analyse

Cette dernière partie de ma thèse prolonge directement la partie précédente puisqu'elle permet de combiner des données de différentes étapes. Elle est aussi basée sur les approches de shrinkage développées dans la première partie de ma thèse. Ici, la méta-analyse est étudiée dans un contexte plus général permettant de combiner des données d'études pour lesquelles une comparaison directe serait impossible mais répondant à une même question biologique. La méta-analyse offre la possibilité d'accroître considérablement la puissance statistique et donne des résultats plus précis. J'ai proposé une approche pour combiner des grandeurs des effets 'modérées' (moderated effect sizes) et l'ai comparée à d'autres approches de méta-analyse (combinaison des grandeurs des effets programmée dans la librairie GeneMeta de Bioconductor, combinaison des p-values par méthode inverse normale). J'ai simulé différentes variabilités inter-études. Bien que la méthode proposée de combiner des grandeurs des effets modérées soit meilleure que les autres approches déjà existantes pour combiner des grandeurs des effets, nous avons montré que les combinaisons de p-values étaient plus performantes que les autres méthodes de méta-analyse en terme de sensibilité. Nous avons aussi étudié des méthodes Bayésiennes pour mieux estimer la variabilité inter-étude, difficile à estimer avec les approches fréquentistes. Ces méthodes Bayésiennes semblent prometteuses pour l'avenir mais nécessitent une mise au point et un temps de calcul très long.

Programmation

J'ai développé des packages R pour chacune de ces questions statistiques. *SMVar* et *metaMA* sont diponibles sur le CRAN, site officiel de R (`http://cran.r-project.org/`).

Publications

Papers:

Marot G., Foulley J.-L., Mayer C.-D. and Jaffrézic F. (2009) Moderated effect size and p-value combinations for microarray meta-analyses. Submitted to Bioinformatics.

Guyonnet B., **Marot G.**, Dacheux J.-L., Lacoste A., Mercat M.-J., Schwob S., Jaffrézic F., Gatti J.-L. (2009) The adult boar testicular and epididymal transcriptomes. Submitted to BMC Genomics.

Marot G. and Mayer C.-D. (2009) Sequential Analysis for Microarray Data Based on Sensitivity and Meta-Analysis. *Statistical Applications in Genetics and Molecular Biology* **8**3(1), Art. 3.

Marot G., Foulley J.-L. and Jaffrézic F. (2009) A structural mixed model to shrink covariance matrices for time-course differential gene expression studies. *Computational Statistics and Data Analysis* **53**(5), p 1630–1638

Jaffrézic F., **Marot G.**, Degrelle S., Hue I. and Foulley, J.-L. (2007) A structural mixed model for variances in differential gene expression studies. *Genetical Research* **89**(1), p. 19-25.

de Koning D-J, Jaffrézic F, Lund M S, Watson M, Channing C, Hulsegge I, Pool M, Buitenhuis B, Hedegaard J, Hornshøj H, Jiang L, Sørensen P, **Marot G**, Delmas C, Lê Cao K-A, SanCristobal M, Baron M D, Malinverni R, Stella A, Brunner R, Seyfert H-M, Jensen K, Mouzaki D, Waddington D, Jiménez-Marín A, Pérez Alegre M, Pérez E, Closset R, Detilleux J, Dovc P, Lavric M, Nie H, Janss L. The EADGENE Microarray Data Analysis Workshop. (2007) *Genetics Selection Evolution*, **39**(6), p. 621-31

Watson M, Pérez Alegre M, Baron M D, Delmas C, Dovc P, Duval M, Foulley J-L, Garrido-Pavón J J, Hulsegge B, Jaffrézic F, Jiménez-Marín Á, Lavriè M, Le Cao K-A, **Marot G**, Mouzaki D, Pool M H, Robert-Granie C, San Cristobal M, Tosser-Klopp G, Waddington D, de Koning D-J. Analysis of a simulated microarray dataset: Comparison of methods for data normalization and detection of differential expression. (2007) *Genetics Selection Evolution*, **39**(6), p. 669-83

Jaffrézic F, de Koning D-J, Boettcher P J, Bonnet A, Buitenhuis B, Closset R, Déjean S, Delmas C, Detilleux J C, Dovc P, Duval M, Foulley J-L, Hedegaard J, Hornshøj H, Hulsegge I B., Janss L, Jensen K, Jiang L, Lavric M, Lê Cao K-A, Lund M S, Malinverni, **Marot G**, Nie H, Petzl W, Pool M H, Robert-Granié C, SanCristobal M, van Schothorst E M., Schuberth H-J, Sørensen P, Stella A, Tosser-Klopp G, Waddington D, Watson M, Yang W, Zerbe H, Seyfert H-M. Analysis of the real EADGENE data set: Comparison of methods and guidelines for data normalization and selection of differentially expressed genes. (2007) *Genetics Selection Evolution*, **39**(6), p. 633-50

Oral presentations (* speaker)

Marot* G., Foulley J.-L, Mayer C.-D., Jaffrézic F. (2009) metaMA : an R package implementing meta-analysis approaches for microarrays. *useR! 2009*, Rennes, France.

Marot* G., Foulley J.-L, Mayer C.-D., Jaffrézic F. (2009) Microarray meta-analysis based on p-value or moderated effect size combinations. *Workshop on Statistical Methods for Post-Genomic Data*, Paris, France.

Marot* G., Jaffrézic F., Foulley J.-L., Mayer C.-D. (2008) Sequential analysis for microarray data based on sensitivity and meta-analysis. *XXIV International Biometric Conference*, Dublin, Ireland.

Marot G., Foulley J.-L., Jaffrézic* F. (2008) A structural mixed model to shrink covariance matrices for time-course differential gene expression studies. *XXIV International Biometric Conference*, Dublin, Ireland.

Marot* G., Foulley J.-L., Jaffrézic F. (2008) Shrinkage of covariance matrices for time-course differential gene expression studies. *Workshop on Statistical Methods for Post-Genomic Data*, Rennes, France.

Marot* G., Mayer C.-D., Foulley J.-L, Jaffrézic F. (2008) Modélisation statistique pour les données d'expression de gènes. *X séminaire des thésards du département de génétique animale*, Toulouse, France.

Guyonnet* B., Dacheux J.-L., Jaffrézic F., Lacoste A., **Marot G.**, Mercat M.-J., Schwob S., Gatti J.-L. (2008) Le transcriptome épididymaire du verrat: étude de la régionalisation. *Journées Recherche Porcine*, 40, 99-104. Paris, France.

Degrelle* S., Hue I., Campion E., Jaffrézic F., **Marot G.**, Everts R., Ducroix-Crépy C., Vignon X., Heyman Y., Yang X., Lewin H., Renard J.-P. (2007) Embryonic and extraembryonic abnormalities in Day-18 cloned bovine conceptuses. *II International Meeting on Mammalian Embryogenomics*, Paris, France.

Jaffrézic* F., **Marot G.**, Degrelle S., Hue I., Foulley J.-L. (2007) Detection of differentially expressed genes: the importance of variance modelling in the test statistics. *II International Meeting on Mammalian Embryogenomics*, Paris, France.

Marot* G., Foulley J-L., Jaffrézic F. (2006) Variance model comparisons for differential gene expression. *EADGENE Data Analysis Workshop*, Tune, Denmark.

Marot G., Foulley J-L., Jaffrézic* F. (2006) Real data analysis with a structural model for variances for differential gene expression. *EADGENE Data Analysis Workshop*, Tune, Denmark.

Posters

Marot* G., Mayer C.-D., Foulley J.-L., Jaffrézic F. (2008) Modélisation statistique pour les données d'expression de gènes *Journées ABIES*, Paris, France.

Guyonnet* B., Dacheux J.-L., Jaffrézic F., Lacoste A., **Marot G.**, Mercat M.-J., Schwob S., Gatti J.-L. (2008) The Adult Boar Testicular and epididymal transcriptome *Society for the Study of Reproduction* Kailua-Kona, USA

Mayer* C.-D. and **Marot G.** (2007) A sequential approach to microarray analyses *IV Annual meeting of NuGO - The European Nutrigenomics Organisation*, Oslo, Norway

Marot* G., Foulley J.-L., Hue I., Mayer C.-D. Jaffrézic F. (2007) Modélisation statistique pour les données d'expression de gènes *IX séminaire des thésards du département de génétique animale*, Jouy en Josas, France

Guyonnet* B., Dacheux J.-L., Jaffrézic F., Lacoste A., **Marot G.**, Mercat M.-J., Gatti J.-L. (2007) Recherche et identification de gènes différentiellement exprimés dans l'épididyme de verrat par une approche transcriptomique. *Journées Recherche Porcine*, 39, 297-298. Paris, France

Guyonnet* B., Dacheux J.-L., Jaffrézic F., Lacoste A., **Marot G.**, Mercat M.-J., Schwob S., Gatti J.-L. (2007) Le transcriptome épididymaire chez le verrat: Etude de la Régionalisation. *Société Française de Biochimie et de Biologie Moléculaire*, Ile des Embiez, France.

Guyonnet* B., Dacheux J.-L., Jaffrézic F., Lacoste A., **Marot G.**, Mercat M.-J., Gatti J.-L. (2006) Analysis of differentially expressed genes along the boar epididymis. *IV International workshop on epididymis*, Châtel-Guyon, France.

Contents

Introduction

Gene expression data represent nowadays an essential stake in human and animal genetics. Hope is allowed that thanks to them and other new technologies, people will be able to model the functioning of cells and genes and consequently, understand the development of individuals, the causes of diseases, etc.

Statistical modelling for microarray data has been a widespread research subject over the past few years and has been dealt within several phD theses, e.g. Delmar (2005), Neuvial (2008), Lê Cao (2008). My thesis brings a contribution to this field, and more precisely to the research for differentially expressed genes. This research often relies on t-tests. While variance modelling from denominators of such test statistics has already been investigated by several authors (Tusher *et al.*, 2001; Kerr *et al.*, 2002; Smyth, 2004; Delmar *et al.*, 2005), the approach I present which takes into account the variance heterogeneity across conditions is new in the microarray analysis context, as well as the extension to variance-covariance matrices. The originality of my PhD especially comes from sequential and meta-analysis. Until now, sequential analysis had never been applied to microarray experiments and meta-analysis is a field which is rapidly expanding with the growing amount of data available. Before explaining in dedicated chapters all these statistical aspects and the developments we propose for microarrays, the next section describes the biological experiments studied in this PhD.

Microarray presentation

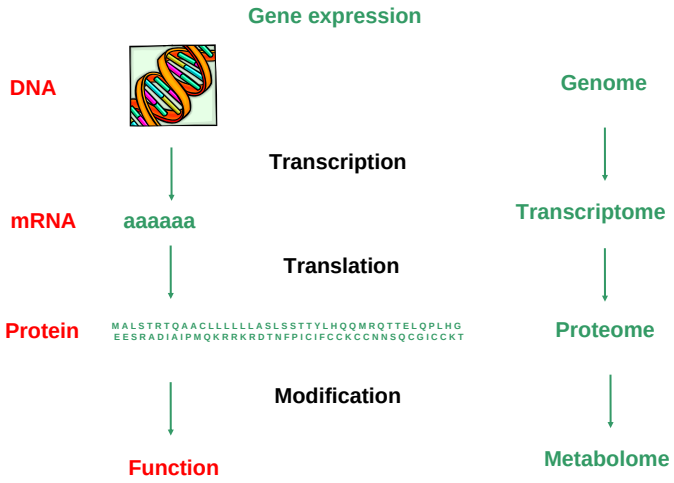
The term 'microarrays' refers to various types of experiments with a broad range of applications (e.g. comparative genomic hybridization[1], gene expression profiling, ChIP on chip, tiling arrays, SNP or alternative splicing detection). In this section, I will only present microarrays which measure gene

[1] A glossary is given in Appendix

expression levels since my PhD is only concerned by this type of experiments.

Since the first gene expression product is a messenger RNA (mRNA), microarrays measuring gene expression are experiments which study the transcriptome. Figure 1 recalls the relationship between genome, transcriptome, proteome, metabolome.

Figure 1: Schematic representation of gene expression study levels

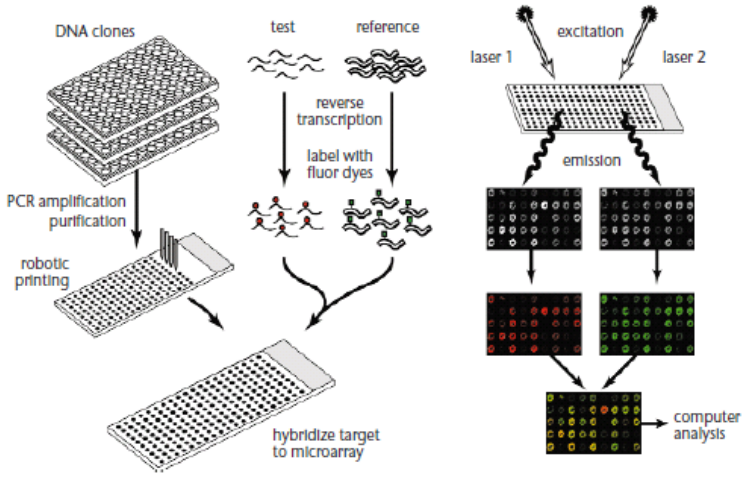


If DNA is not transcribed in mRNA then there is no possible translation and absence of protein often means absence of associated function. Variations in transcriptome can be physiological (during growth, daily situations as adaptation to effort) or pathological (diseases can be caused by abnormal gene regulations). As for a given individual, genes are identical from a tissue to another one, a cell to another one, it is important to look at relative gene expression levels to distinguish functions of cells and understand the mechanism of regulation. In the same way, the differences between the gene expression levels from healthy patients and the ones from ill patients are explanations of disregulations which could be the cause of the studied disease. That is why people are interested in studying the transcriptome. Quantification of gene expression levels can be performed via microarrays. Thus, one application of these experiments is to look for differentially expressed genes, that is to say genes which would be up or down regulated between two 'conditions'. Note that the term 'conditions' is often used whenever the studied problem is formalised statistically. It refers to the different origins of the harvested samples (e.g. normal/tumoral samples). Microarrays can

measure the expression of thousands of genes simultaneously.

The principle of microarrays is based on a property of DNA and RNA: the hybridization (recognition and interaction of two complementary sequences of RNA or DNA). The principle of these experiments is shown in Figure 2 (Duggan *et al.*, 1999).

Figure 2: Principle of microarray experiments



Chips are solid supports (e.g. glass slide, nylon membrane) on which what is called 'probes' are spotted. Probes can be oligonucleotides (synthesized directly on the array surface for oligonucleotides arrays or prior to the deposition on the array for spotted arrays), cDNA clones (much longer than oligos, they can be thousands of nucleotides long). Many steps (not detailed here for simplicity) are satisfied to have good quality chips. What is common between all the different types of probes is that they only have one strand and will be able to capture 'targets'. Targets are the mRNA harvested from the studied cells. They are marked either by fluorescence or radioactivity so that when they hybridize with probes, the quantity can easily be measured by scanner. The example given in Figure 2 represents a two-color array. In this case, cells from the two conditions (e.g. control/target cells) have been tagged by two different fluorophores: Cyanine 3 (Cy3) and Cyanine 5 (Cy5). The scanner generates an image with levels of grey depending on the fluorochrome intensity read. A representation with false colors is often used. These colors vary from green, which caracterizes samples marked with Cy3 to red for Cy5. Thus, if target cells have been fluorophored with Cy3 then a green spot underlines an over-expression of the corresponding gene in these

cells compared to the control ones. The spot is yellow when no differential expression is observed. In addition to differences observed on supports and probes, the type of microarrays also depends on the number of channels considered. In my PhD report, there are two channels for two-color arrays and one channel for nylon membrane chips. The data that I have downloaded on public repositories mostly come from one channel microarrays since they most often correspond to commercial Affymetrix chips (oligonucleotide arrays).

Statistical problem

Microarrays generate a large quantity of data (thousands of genes studied at the same time) with very few samples (less than ten per condition most of the time). In the following of this report, the terminology adopted for 'samples' is 'replicates'. Biological replicates are distinguished from technical replicates. Biological replicates are samples coming from different individuals with caracteristics from the same condition while technical replicates come from the same individuals and should give the same results. When not precised, the term 'replicates' refers to biological replicates. From the statistical point of view, this is a typical '$n \ll p$' problem with a much larger number of variables (p: number of genes) than experimental units (n: number of individuals/replicates). Until now, when looking for differentially expressed genes, this problem has mostly been considered from the multiple testing side. Indeed, since many tests (one per gene) are performed simultaneously, there are many false positives (genes which are declared differentially expressed while they are not). Solutions to this problem have been brought by Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001), Storey and Tibshirani (2003), McLachlan *et al.* (2006), Robin *et al.* (2007) and others. In the following, I will mostly use the Benjamini Hochberg correction (Benjamini and Hochberg, 1995) to adjust p-values for multiple testing and thus control the False Discovery Rate (FDR), the expected proportion of false positives. Other authors have contributed to the microarray field developing methods to reduce high dimensionality, which is particularly needed when people perform classifications or when the aim of the experiment is to predict the belonging of a sample to one condition from a subset of genes (Lê Cao, 2008; Mary-Huard, 2006).

The purpose of my PhD is a bit different from the previous ones. I most often consider high dimensionality as an advantage and try to gain information from it in order to palliate the small sample size problem. All my work is concerned with the research of differentially expressed genes and the aim is to increase sensitivity, that is to say the expected proportion of true positives

(both declared significant and truly differentially expressed genes) among the truly differentially expressed genes. The first part of my PhD concentrates on variance-covariance modelling. Sensitivity is increased thanks to shrinkage approaches. After that, sequential analysis is developed in order to save samples as soon as significant results are obtained. The notion of stopping rule is explained in the corresponding part. It is sometimes achieved by looking at sensitivity but we will see the difficulties to estimate this criterion. Finally, the last part of my PhD takes advantage of shrinkage approaches presented in the first part to extend meta-analysis approaches introduced in the second part. The aim is to increase sensitivity by gathering several studies where only few samples are involved in each of them.

The following of this report will only develop my specific contributions to the statistical analysis of microarray data. For a classical analysis of microarray data, I would advise people to read books like Speed (2003), Wit and McClure (2004), Mary-Huard *et al.* (2006). In particular, I always assume that data have already been normalised. This point will not be developed since I did not bring any new contribution but it is really an important step of microarray analysis. For two-color arrays, this normalisation most often consists in at least correcting for dye effects (by loess correction (Yang *et al.*, 2002) for example) and block effects induced by prints of the robot used for spotting. For many types of microarrays, intensity levels are log-transformed during the normalisation process, which explains that comparing the ratio between two conditions can be achieved by a simple difference between gene expression normalised levels. Affymetrix arrays need a normalisation which takes into account the special design (Perfect Match/Mismatch probe strategy) of oligonucleotide arrays. In all cases, only one normalised value is kept for each replicate in each condition for each gene. This is what is called later the gene expression level. When data are paired (for example for two color arrays where the two conditions are on the same chip), it is preferable to directly consider the log-ratio that is to say the difference between the gene expression levels of the two conditions.

Part I

Variance-covariance modelling for differential gene expression studies

Transcriptomic studies using microarray technology have become a standard tool in life sciences over the past decade. However, the cost of these experiments remains high, which often results in a lack of samples available. When very few samples are considered and the analysis is performed gene-by-gene, statistical tests lack of power; in this context, it means that very few differentially expressed genes can be detected. One alternative to that is to assume a common variance between all genes. However, it often results in an increase of false positives. To overcome these problems, we investigated shrinkage approaches.

Shrinkage is a key-word, which will be very often used in this report since both variance-covariance modelling and meta-analysis rely on shrinkage approaches. Historically, the shrinkage concept comes from James and Stein (1961) who proposed a nonlinear estimator which outperforms the ordinary least squares technique. The shrinkage procedure improves the efficiency of the resulting estimator with respect to Mean Square Error. Actually, the word 'shrinkage' is used whenever an estimator is a compromise between two estimators. In a general situation, the shrinkage estimator $\tilde{\boldsymbol{\theta}}_{\boldsymbol{g}}$ can be written as a function of a gene-by-gene estimator $\widehat{\boldsymbol{\theta}_{\boldsymbol{g}}}$ and a common estimator of the whole population $\widehat{\boldsymbol{\theta}_{\boldsymbol{c}}}$:

$$\tilde{\boldsymbol{\theta}}_{\boldsymbol{g}} = \widehat{\boldsymbol{\theta}_{\boldsymbol{c}}} + b(\widehat{\boldsymbol{\theta}_{\boldsymbol{g}}} - \widehat{\boldsymbol{\theta}_{\boldsymbol{c}}}) \tag{2}$$

where b is the shrinkage factor. When $b = 1$, $\tilde{\boldsymbol{\theta}}_{\boldsymbol{g}} = \widehat{\boldsymbol{\theta}_{\boldsymbol{g}}}$ (gene-by-gene empirical estimator). When $b = 0$, $\tilde{\boldsymbol{\theta}}_{\boldsymbol{g}} = \widehat{\boldsymbol{\theta}_{\boldsymbol{c}}}$ (common estimator). Shrinkage approaches considerably decrease the number of parameters to estimate while still keeping a certain flexibility with one value per gene.

Variance modelling plays an important role in differential gene expression studies. When very few samples are considered and the analysis is performed gene-by-gene, statistical tests lack of power; in this context, it means that very few differentially expressed genes can be detected. One alternative to that is to assume a common variance between all genes. However, it often results in an increase of false positives. During my end of study ENSAI internship, I implemented a novel shrinkage approach based on a structural mixed model for variances. In this case, θ in equation 2 corresponds to the log of the variances and is written as a mixed model with a fixed condition effect and a random gene effect. The shrinkage factor is estimated via an empirical Bayesian approach. This approach is presented at the beginning of Chapter 1. At the beginning of my PhD, I extended this variance modelling to covariance modelling, which is particularly important for time-course studies when measures are repeated on the same individuals. With F. Jaffrézic and

J.-L. Foulley (INRA), we proposed to apply a shrinkage approach based on a structural mixed model via an eigenvalue and a Cholesky decomposition of the variance-covariance matrices. Shrinkage estimators were derived at three levels i) the eigenvalues, ii) the innovation variances, iii) both the variances and correlation parameters of a gene-by-gene covariance matrix. This natural extension to variance-covariance modelling is provided in Chapter 2.

Chapter 1

Modelling of variances

Differential gene expression studies often rely on t-tests or F-tests. Due to the high number of parameters involved in microarray experiments and the small number of samples available, variance-covariance modelling plays an important role. Following the work of Tusher *et al.* (2001), Kerr *et al.* (2002), Smyth (2004), Delmar *et al.* (2005) on variance modelling, we proposed a 'shrinkage' method based on a structural model.

1.1 A sructural mixed model for variances in differential gene expression studies

Genet. Res., Camb. (2007), **89**, *pp.* 19–25. © 2007 Cambridge University Press
doi:10.1017/S0016672307008646 Printed in the United Kingdom

A structural mixed model for variances in differential gene expression studies

FLORENCE JAFFRÉZIC[1]*, GUILLEMETTE MAROT[1], SÉVERINE DEGRELLE[2], ISABELLE HUE[2] AND JEAN-LOUIS FOULLEY[1]
[1] *INRA, UR337 Station de Génétique Quantitative et Appliquée, Jouy-en-Josas 78350, France*
[2] *INRA, UMR 1198; ENVA; CNRS, FRE 2857, Biologie du Développement et Reproduction, Jouy-en-Josas 78350, France*

(*Received 12 March 2007 and in revised form 6 April 2007*)

Genet. Res., Camb. (2007), **89**, *pp.* 19–25. © 2007 Cambridge University Press
doi:10.1017/S0016672307008646 Printed in the United Kingdom

A structural mixed model for variances in differential gene expression studies

FLORENCE JAFFRÉZIC[1]*, GUILLEMETTE MAROT[1], SÉVERINE DEGRELLE[2], ISABELLE HUE[2] AND JEAN-LOUIS FOULLEY[1]
[1] *INRA, UR337 Station de Génétique Quantitative et Appliquée, Jouy-en-Josas 78350, France*
[2] *INRA, UMR 1198; ENVA; CNRS, FRE 2857, Biologie du Développement et Reproduction, Jouy-en-Josas 78350, France*

(*Received 12 March 2007 and in revised form 6 April 2007*)

Summary

The importance of variance modelling is now widely known for the analysis of microarray data. In particular the power and accuracy of statistical tests for differential gene expressions are highly dependent on variance modelling. The aim of this paper is to use a structural model on the variances, which includes a condition effect and a random gene effect, and to propose a simple estimation procedure for these parameters by working on the empirical variances. The proposed variance model was compared with various methods on both real and simulated data. It proved to be more powerful than the gene-by-gene analysis and more robust to the number of false positives than the homogeneous variance model. It performed well compared with recently proposed approaches such as SAM and VarMixt even for a small number of replicates, and performed similarly to Limma. The main advantage of the structural model is that, thanks to the use of a linear mixed model on the logarithm of the variances, various factors of variation can easily be incorporated in the model, which is not the case for previously proposed empirical Bayes methods. It is also very fast to compute and is adapted to the comparison of more than two conditions.

1. Introduction

Detection of differentially expressed genes relies on statistical tests, typically *t*-tests. A key and critical aspect of these tests is the modelling of the residual variances. The most commonly used approach is to test for differential gene expression one gene at a time. This approach has, in general, low power due to the lack of information on each individual gene (Callow *et al*., 2000). On the other hand, assuming that all the variances are equal and using a common variance estimator can increase the power (Kerr *et al*., 2000) but generates a high rate of false positives when the assumption of homoskedasticity is not true (Cui *et al*., 2005). A number of papers have been devoted to the problem of choosing a suitable variance model for microarray data. In the SAM *t*-test (Tusher *et al*., 2001) a small constant is added to the gene-specific variance estimates in order to stabilize the small variances. Kerr *et al*. (2002) proposed an intensity-dependent variance model where the gene-specific residual variances are modelled as a non-parametric function of the log-intensity. Delmar *et al*. (2005*a*) proposed a mixture model on the gene-variance distributions to identify clusters of genes with equal variances. Cui *et al*. (2005) presented a shrinkage estimator of variance components, using the James–Stein shrinkage concept. Several authors have also proposed hierarchical Bayesian methods, including Lewin *et al*. (2006), Newton *et al*. (2001), Baldi & Long (2001), Lönnstedt & Speed (2002), Wright & Simon (2003), Smyth (2004) and Feng *et al*. (2006).

The aim of this paper is to propose a simple and biologically interpretable model for the variances. The idea is to consider a structural model (Foulley *et al*., 1992) which includes a condition effect and a random gene effect. This model will allow estimation of gene-specific residual variances that will take into account information from all the genes in the data set in a simple and parsimonious way. Two estimation procedures are considered in this paper to estimate the variance parameters: a stochastic approach based

* Corresponding author. e-mail: florence.jaffrezic@jouy.inra.fr

on MCMC techniques and a simple approximate method.

In a simulation study, our method was compared with five other approaches for variance modelling: gene-specific variances, common variance model, SAM (Tusher *et al.*, 2001), VarMixt (Delmar *et al.*, 2005*b*) and Limma (Smyth, 2004). The proposed structural model was also applied to a real functional genomics study on bovine embryos before implantation to find differentially expressed genes according to the reproduction mode, and to a microarray experiment to study the response of the mouse spleen to *in vivo* whole body irradiation.

2. Materials and methods

(i) *Hierarchical model*

Let y_{ijk} be the expression level for gene i ($i=1, \ldots, N$), replicate j ($j=1, \ldots, n_i$) and condition k ($k=1, \ldots, K$). Data are assumed to have been previously normalized. Observations y_{ijk} are modelled with the simple linear model (Delmar *et al.*, 2005*a*):

$$y_{ijk} = m_{ik} + e_{ijk}. \tag{1}$$

The residual terms e_{ijk} are assumed to be independent and normally distributed with mean zero and a variance which can vary both by gene and condition: $e_{ijk} \sim \mathcal{N}(0, \sigma^2_{ik})$.

Estimating one residual variance for each gene within each condition is often not possible due to the lack of replications within each interaction cell. The second step of the proposed hierarchical modelling is therefore to consider a model on the variances that will retain flexibility while keeping the number of parameters reasonably low. As suggested by Foulley *et al.* (1992), a structural model is therefore assumed on the logarithm of the residual variances:

$$\ln(\sigma^2_{ik}) = \mu_k + \delta_{ik}, \tag{2}$$

where μ_k is a condition effect (assumed fixed) and δ_{ik} is the gene effect in condition k. Here we will assume that the gene effects are independent and normally distributed with mean zero and variance τ^2_k, i.e. $\delta_{ik} \sim \mathcal{N}(0, \tau^2_k)$. Considering the gene effects as random allows us to take into account this source of variation parsimoniously and leads, as shown later, to a shrunk estimator of the variance.

(ii) *Simple estimation procedure*

Analytical forms of the likelihood function are difficult to obtain in the model presented above, and estimation of the parameters in such a structural model for the variances usually requires the use of stochastic estimation procedures based on MCMC methods. Lewin *et al.* (2006), for example, proposed using Gibbs sampling and estimated the parameters in a Bayesian framework. These stochastic estimation procedures are, however, quite time-consuming due to the large number of simulations required to obtain accurate estimates of the parameters.

Here we propose a simple and efficient approximate method to obtain estimates of the parameters in the structural model for the variances. These estimates were compared with those obtained with Gibbs sampling using the software WINBUGS (Spiegelhalter *et al.*, 2004). The idea of the proposed estimation procedure is to base inference of the variance parameters on the empirical variances.

For each gene i, let s^2_{ik} be the empirical variance defined as

$$s^2_{ik} = \frac{1}{n_{ik}-1} \sum_{j=1}^{n_{ik}} (y_{ijk} - y_{ik.})^2, \tag{3}$$

where y_{ijk} represents the expression level for replicate j of gene i in condition k. Let $y_{ik.}$ be the average expression level for gene i over all replicates in condition k: $y_{ik.} = \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} y_{ijk}$. For the proposed estimation procedure, the structural model is assumed on the logarithm of the empirical variances:

$$\ln(s^2_{ik}) = \mu_k + \delta_{ik} + \varepsilon_{ik}, \tag{4}$$

where ε_{ik} is a sampling error due to the estimation of the true variances σ^2_{ik} by the empirical variances s^2_{ik}. Residuals ε_{ik} are assumed independent and normally distributed with mean zero and variance ω^2_{ik}: $\varepsilon_{ik} \sim \mathcal{N}(0, \omega^2_{ik})$. According to the asymptotic theory (Layard, 1973), the sampling variances ω^2_{ik} can be estimated by $\omega^2_{ik} = 2/d_{ik}$, where d_{ik} corresponds to the degrees of freedom for gene i in condition k. Usually $d_{ik} = n_{ik} - 1$, where n_{ik} represents the number of replicates for gene i in condition k. As previously, δ_{ik} is assumed to be a random gene effect in condition k: $\delta_{ik} \sim \mathcal{N}(0, \tau^2_k)$, and μ_k is a fixed effect which represents the condition effect. Both parameters τ^2_k and μ_k can be estimated by classical linear mixed model estimation procedures.

Due to the use of normal conjugate distributions – $\ln s^2_{ik} | \ln \sigma^2_{ik} \sim \mathcal{N}(\ln \sigma^2_{ik}, \omega^2_{ik})$ and $\ln \sigma^2_{ik} \sim \mathcal{N}(\mu_k, \tau^2_k)$ – it follows that the best predictor of $\ln \sigma^2_{ik}$ is

$$\widehat{\ln \sigma^2_{ik}} = \mu_k + \lambda_{ik}(\ln s^2_{ik} - \mu_k), \tag{5}$$

where $\lambda_{ik} = \tau^2_k / (\tau^2_k + \omega^2_{ik})$ is a shrinkage factor of $\ln \sigma^2_{ik}$ towards μ_k. When parameters τ^2_k tend to zero, we obtain a pooled estimator and a common variance for all genes within each condition: $\hat{\mu}_k = \sum_i (d_{ik} \ln s^2_{ik}) / \sum_i d_{ik}$. On the other hand, if parameters τ^2_k tend to infinity, the shrinkage factors λ_{ik} become 1. There is no shrinkage, and one variance is estimated for each gene in each condition as: $\widehat{\ln \sigma^2_{ik}} = \ln s^2_{ik}$.

(iii) *Degrees of freedom of the T statistic*

To test whether gene i is differentially expressed between condition k and condition l the test statistic is

$$t_{i,kl} = \frac{m_{ik} - m_{il}}{\sqrt{\hat{\sigma}^2_{ik}/n_{ik} + \hat{\sigma}^2_{il}/n_{il}}}, \quad (6)$$

where $\hat{\sigma}^2_{ik}$ and $\hat{\sigma}^2_{il}$ are estimations under the proposed structural model presented above. The exact distribution of this test statistic under the null hypothesis is unknown and determination of the p values can be obtained by permutations. As pointed out by Cui *et al.* (2005) permutations are, however, very time-consuming, especially when a large number of genes are analysed. To obtain a fast and efficient procedure, we therefore propose considering an approximate Student distribution. In fact, under the structural model, the test statistic corresponds to the so-called Welch's statistic which follows approximately a Student distribution (Moser & Stevens, 1992) with ν_i degrees of freedom. For each gene, we propose calculating the degrees of freedom of the T statistic by the classically used Satterthwaite's method, as follows:

$$\nu_i = \frac{2(\hat{\sigma}^2_{ik} + \hat{\sigma}^2_{il})^2}{\mathrm{Var}(\hat{\sigma}^2_{ik}) + \mathrm{Var}(\hat{\sigma}^2_{il})}, \quad (7)$$

where $\hat{\sigma}^2_{ik}$ and $\hat{\sigma}^2_{il}$ are the variance parameter estimations obtained with the structural model and the variances of these estimations can be calculated as: $\mathrm{Var}(\hat{\sigma}^2_{ik}) = (\hat{\sigma}^2_{ik})^2 \mathrm{Var}(\widehat{\ln \sigma^2_{ik}})$, where $\mathrm{Var}(\widehat{\ln \sigma^2_{ik}}) \approx (1/\tau^2_k + d_{ik}/2)^{-1}$ with $d_{ik} = (n_{ik} - 1)$ for condition k.

An R function 'SMVar' implementing the structural model for the detection of differentially expressed genes is available upon request from the first or second author.

3. Application

The proposed structural model was applied here to two sets of real data to find differentially expressed genes in bovine embryos according to the reproductive mode and in mice to study the spleen response to irradiation.

(i) *Reproductive mode in bovine embryos*

(a) *Presentation of the data.* This variance modelling was applied to a functional genomics study on bovine embryos before implantation. The experimental protocol is described in detail by Degrelle (2006). The aim of this study was to find differentially expressed genes in the embryos according to the reproductive mode. Three reproductive modes were investigated: artificial insemination (AI), *in vitro* fertilization (IVF) and cloning (somatic cell nuclear transfer, SCNT). Three different lines of clones were studied. They were established from ear skin biopsies of three Holstein heifers. In total, 10 Holstein embryos were available for AI, IVF and each of the three lines of clones. In total, 10 214 unique cDNA were spotted onto Nylon N+ membranes (Amersham Biosciences) at the CRB GADIE platform (INRA, Jouy-en-Josas). The bovine 10K array will be fully described in a forthcoming paper (Degrelle *et al.*, unpublished). For each embryo ($n = 50$), RNA was isolated, amplified (MessageAmp aRNA Kit, Ambion) and hybridized onto the array. The membranes were exposed to phosphor screens for 7 days. The hybridization signals were quantified using Imagene 5.5 software (BioDiscovery) on the PICT platform (INRA, Jouy-en-Josas). Gene expression data were $\log_2$ transformed. Data were centred by membrane and by gene. No further normalization was needed on this bovine data set.

(b) *Variance parameter estimations.* For the structural model, the list of differentially expressed genes found with the approximated estimation method was compared with the list obtained with the exact MCMC estimations using Gibbs sampling with WINBUGS software (Spiegelhalter *et al.*, 2004). As the posterior distributions of the variance parameters were highly asymmetrical, we chose the posterior mode with a uniform prior on the standard deviations (Gelman, 2005) as a point estimate of the variance parameters, which is close to the REML estimation of the variance parameters. The structural model was compared with the mixture model approaches proposed by Delmar *et al.* (2005*b*): VM and VM2. In VM2, each gene is assigned to one of the groups of homogeneous variance determined by the mixture model. VM is more flexible as it does a partial assignment of genes to variance groups, taking into account the probabilities of belonging to each group. Classical methods such as Limma (Smyth, 2004), SAM (Tusher *et al.*, 2001), gene-by-gene analysis and the homogeneous variance model were also applied to this data set. To make each method comparable, a Benjamini & Hochberg (1995) correction (BH correction) was performed on the raw p values to correct for multiple tests.

The proposed structural model is similar in spirit to that of Baldi & Long (2001), except that the use of log-normal distributions instead of Gamma gives the possibility of directly estimating the shrinkage parameter, which is a crucial parameter for the variance estimations, whereas it has to be specified *a priori* by the user in Cyber-T. Moreover, the structural model allows the easy incorporation of factors of variation other than the gene and condition effects. Analyses performed here will therefore not be

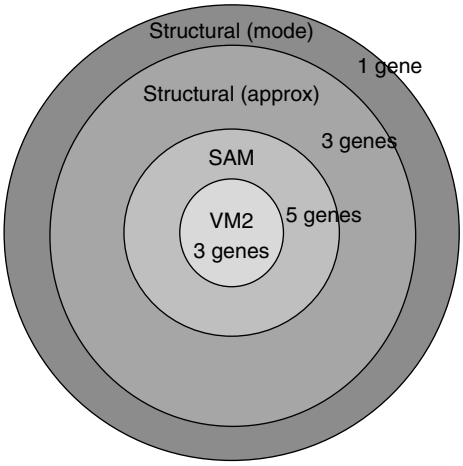


Fig. 1. Venn diagram for the differentially expressed genes detected at a 10% BH threshold in the real bovine data set with four methods: structural model using the posterior mode in the Gibbs sampling estimations, structural model with the approximate method, SAM and VM2.

compared with the method presented by Baldi & Long (2001).

(c) *Results*. The Venn diagram presented in Fig. 1 illustrates the list of differentially expressed genes found with the different methods at a 10% BH threshold. With the structural model and the proposed approximate estimation method, 11 genes were found which were all included in the 12 genes found with Gibbs sampling estimations using posterior mode estimates.

The SAM approach detected 8 genes at 10% which were all included in the 11 genes detected with the structural model. In this analysis, VM and VM2 were found to lack power as no gene was detected with VM and only 3 genes were detected with VM2. This may be due to the fact that the VarMixt methods were designed for comparing only two conditions as the variance of the gene expression difference is modelled, whereas in this example five conditions were compared. In contrast, the structural approach models the variances in each condition and can therefore readily be applied to the comparison of more than two conditions. At a 30% BH threshold, 9 genes were detected with VM and 6 with VM2. All of them were included in the 11 genes detected with the structural model. Similarly, Limma detected only 1 differentially expressed gene at a BH threshold of 10% as well as 30%.

The homogeneous variance model found far more genes than other methods, but a histogram of the p values showed that the assumption of a common variance is not appropriate for these data, as shown in Fig. 2. In fact, the distribution of the p values was not uniform under the null hypothesis. It is therefore expected that a large proportion of the detected genes are false positives.

(ii) *Mouse spleen data*

(a) *Presentation of the data*. These data were presented and analysed by Delmar *et al*. (2005*a*), and are publicly available in the R VarMixt package (Delmar *et al*., 2005*b*). The goal of this experiment was to study the response of the mouse spleen to *in vivo* whole-body irradiation. Experimental data were generated with two-colour complementary DNA microarray assays comparing the spleen of irradiated (treated) and normal (control) mice. The data consist of three dye-swaps. The 'treated' samples were obtained from three independent mice (one mouse per swap) 3 hours after irradiation at 1 Gy. The 'control' sample was obtained from pooling several normal mice. The same control sample was used in all the hybridization experiments. There are 4360 genes in each array. Composition of the arrays is described in Preisser *et al*. (2004). Data were previously normalized as described by Delmar *et al*. (2005*a*).

(b) *Results*. Three methods have been applied to find differentially expressed genes in these data, namely Limma (Smyth, 2004), VM (Delmar *et al*., 2005*b*) and the structural model proposed here. The Benjamini & Hochberg (1995) procedure at a 5% threshold was used to correct for multiple tests. In total 112 genes were detected with Limma, 113 with VM and 125 with the structural model. Among them, 104 genes were found by all three methods, as shown in the Venn diagram in Fig. 3.

4. Simulation study

A simulation study was performed to compare the proposed structural model with the variance modelling implemented in Limma (Smyth, 2004), SAM (Tusher *et al*., 2001) and VarMixt (Delmar *et al*., 2005*b*), as well as with the simple gene-by-gene analyis and homogenenous variance model. In the first simulation, paired data were studied from the 'two-colour' experiment in mice presented by Delmar *et al*. (2005*a*) and analysed in the previous section. The second simulation study is based on the real bovine data presented above; these are therefore unpaired data. For each of the methods, a BH correction was performed on the raw p values to account for multiple tests.

(i) *Simulation 1*

(a) *Data*. The first simulation was performed with the same parameters as used by Delmar *et al*.

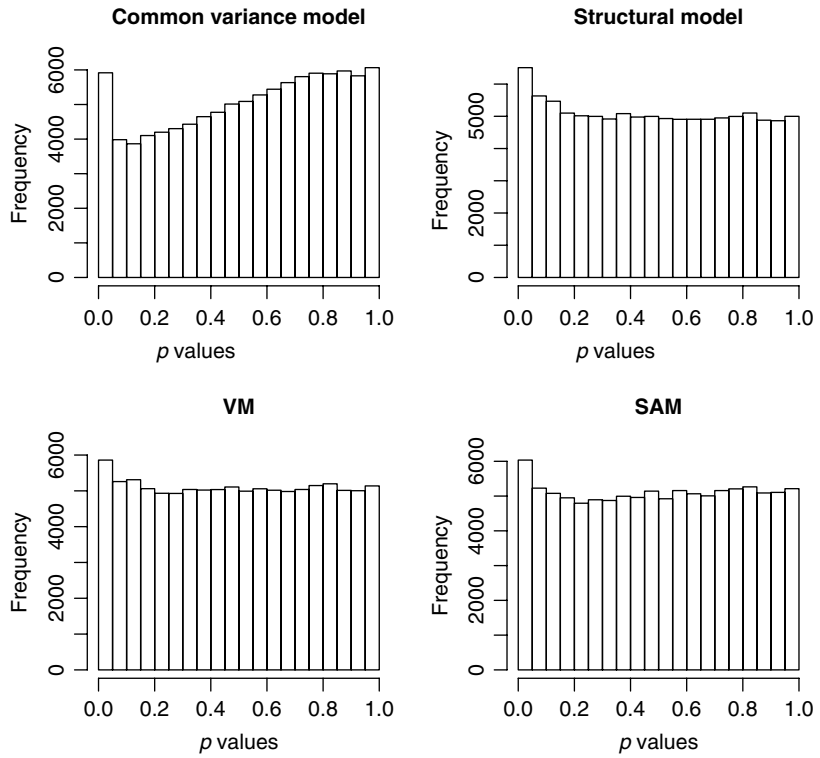


Fig. 2. Histogram of the raw p values for the real bovine data analysis with four different models: the common variance model, the proposed structural model, VM and SAM.

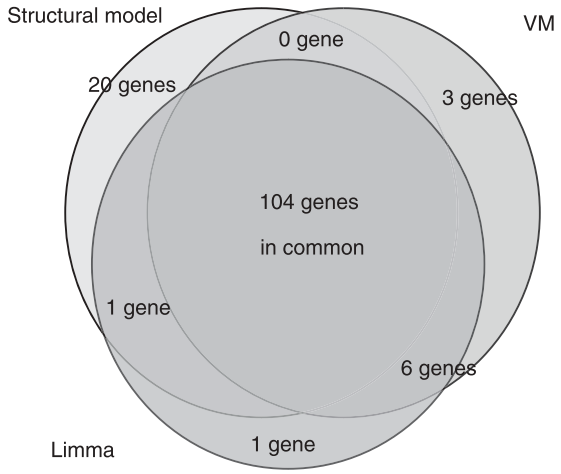


Fig. 3. Venn diagram for the differentially expressed genes detected at a 5% BH threshold in the real mouse data set with three methods: structural model, VM and Limma.

(2005*a*), which were obtained from the real mouse data set analysed above. The simulated data had 4360 genes which were assumed normally distributed. One per cent of the genes were simulated with a non-zero mean log-ratio. These 43 genes were simulated with a mean log-ratio ranging uniformly from 0.25 to 0.9. Gene variances were estimated from the real data by a gene-by-gene analysis and were randomly assigned to the differentially expressed genes in each simulated data set as in Delmar *et al*. (2005*a*).

(b) *Model fitting*. In this data set, each array is hybridized with both a control and a treated sample. Therefore, for each gene the two observations from the same array were treated as paired data. Each model was fitted on the logarithm of the ratio of observed intensity (log-ratio). Let y_{ij} be the log-ratio for gene i in replicate j. It is modelled by

$$y_{ij} = m_i + e_{ij}, \tag{8}$$

where $e_{ij} \sim \mathcal{N}(0, \sigma_i^2)$. For the structural model, the residual variances are now modelled as: $\ln \sigma_i^2 = \mu + \delta_i$, where $\delta_i \sim \mathcal{N}(0, \tau^2)$. For these paired data, the measure of differential expression for gene i between the two conditions is now defined as the mean log-ratio for gene i:

$$\Delta_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}. \tag{9}$$

In this first simulation study, the paired Limma, VM, VM2, SAM, gene-by-gene and homoskedastic models were also used. The results were averaged over 100 simulated data sets and are presented in Table 1 for a 5% BH threshold.

(c) *Results*. It was found that for relatively large numbers of replicates (eight or more), all four methods (Limma, SAM, VarMixt and structural model) perform quite well. The homogeneous model,

Table 1. *Results of the simulations based on the mouse paired data set at a 5% BH threshold*

	No. of replicates[a]		
	5	8	10
No. of true positives			
Structural model	29.4 (3.1)	39.8 (1.6)	41.4 (1.2)
VM	33.7 (2.4)	40.2 (1.4)	41.6 (1.1)
VM2	32.5 (2.7)	39.7 (1.7)	41.3 (1.3)
SAM	0.0 (0.0)	39.9 (1.7)	40.9 (1.5)
Limma	33.0 (2.5)	40.0 (1.6)	41.4 (1.2)
Gene-specific	13.8 (4.2)	37.1 (2.2)	39.9 (1.7)
Homoskedastic	39.9 (1.4)	42.4 (0.8)	42.8 (0.5)
No. of false positives			
Structural model	1.7 (1.5)	2.0 (1.3)	2.2 (1.8)
VM	2.0 (1.6)	2.3 (1.6)	2.4 (1.8)
VM2	2.0 (1.7)	2.2 (1.4)	2.1 (1.7)
SAM	0.0 (0.0)	1.7 (1.5)	2.1 (1.7)
Limma	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
Gene-specific	0.8 (1.0)	1.9 (1.2)	2.4 (1.8)
Homoskedastic	49.8 (7.9)	50.0 (7.9)	51.7 (8.7)

Values are the mean (SD) over 100 simulations.
[a] Replicates correspond to the number of measurements for each gene within each condition.

Table 2. *Results of the simulations based on a subset of the bovine reproductive data set at a 10% BH threshold*

	No. of replicates[a]		
	5	8	10
No. of true positives			
Structural model	18.7 (6.6)	53.9 (5.6)	56.2 (4.6)
VM	10.4 (5.5)	47.7 (5.2)	60.2 (4.7)
VM2	6.4 (5.1)	43.1 (5.5)	58.2 (4.8)
SAM	11.5 (5.5)	50.6 (5.5)	63.3 (4.8)
Limma	17.8 (6.0)	49.3 (5.0)	61.2 (4.4)
Gene-specific	6.2 (4.1)	39.2 (5.0)	54.1 (4.4)
Homoskedastic	37.5 (4.43)	63.2 (4.4)	73.5 (3.9)
No. of false positives			
Structural model	6.2 (4.2)	14.8 (5.3)	16.8 (5.1)
VM	2.0 (1.9)	8.7 (3.7)	11.6 (4.0)
VM2	1.3 (1.7)	7.3 (3.6)	10.3 (3.7)
SAM	2.3 (2.2)	11.2 (4.5)	13.7 (4.6)
Limma	5.0 (3.4)	14.0 (4.9)	16.8 (5.0)
Gene-specific	1.2 (1.5)	7.9 (3.6)	11.3 (4.3)
Homoskedastic	89.1 (11.7)	102.1 (10.6)	104.5 (11.2)

Values are the mean (SD) over 100 simulations.
[a] Replicates correspond to the number of measurements for each gene within each condition.

however, had a very large rate of false positives, which shows that this assumption is here highly unrealistic. The structural model still performs quite well for fewer replicates (five replicates), even with the approximate estimation method based on empirical variances. As already observed by Delmar *et al.* (2005*a*), however, the paired SAM method performs very poorly for five replicates as it detects no differentially expressed genes. As expected, the gene-by-gene analysis showed a lack of power with this small number of replicates and the homogeneous variance model still had a large number of false positives.

(ii) *Simulation 2*

(a) *Data*. This second simulation was based on the parameters estimated on the real data set presented above on bovine embryos. Due to the required computing time, only three conditions among the five were considered and 5000 genes among the 10 214. Among them, 100 genes were simulated to be differentially expressed for one of the conditions compared with the two others. For these genes, the mean log-ratio was simulated according to a Gamma (10.8,0.07). These parameters were determined from the real data set. Gene variances used for the simulations were estimated from the real data by a gene-by-gene analysis and were randomly distributed within the set of differentially expressed genes for each simulation. In this study, only unpaired methods were used as the real data came from a membrane experiment and not a 'two-colour' experiment. Results were averaged over 100 simulated data sets and are presented in Table 2 for a 10% BH threshold.

(b) *Results*. In the case of the comparison of more than two conditions, as already observed in the real data analysis, the structural model had more power than VM and SAM, especially in the case of a small number of replicates (five replicates here). In fact, 19 true positives were detected on average with the structural model at a 10% BH threshold, whereas fewer than seven genes were detected with VM2, fewer than 11 with VM and 12 with SAM. This is due to the fact that the structural approach models directly the variance of each gene within each condition, whereas the VarMixt methods model the variance of the difference in gene expression in two conditions. On the other hand, the Limma approach also works quite well in this case with 18 true positives detected. In the case of 10 replicates the same pattern is oberved, although the differences between methods are slightly smaller than for five replicates.

5. Discussion

The first simulation study showed that the proposed structural model for paired data performed similarly to the VarMixt approach. The paired SAM method, however, showed a considerable lack of power in this

analysis when the number of replicates was small. As expected, the structural model clearly outperformed the homogeneous variance model and the gene-by-gene analysis. The proposed approximate estimation procedure, based on empirical variances, still performed well for a small number of replicates (five replicates).

In the first real data analysis and the second simulation study, the structural model was found to be more powerful than VM, VM2 and SAM. This was due to the fact that more than two conditions were compared whereas VM and VM2 were initially developed for the comparison of only two conditions. In fact, the mixture model is based directly on the variance of the gene expression difference instead of modelling the variance in each condition.

The structural model was found here to perform similarly to the Limma approach (Smyth, 2004). The main advantage of the structural model is, however, that the use of a linear mixed model on the log of the variances provides a larger modelling flexibility. In fact, here a condition and gene effects were included in the model, but it could easily be extended to other mixed models including, for example, a sex effect or even functions of time. This is much more difficult to achieve when considering an inverse chi-square distribution on the variances, as proposed by Smyth (2004).

References

Baldi, P. & Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **85**, 289–300.

Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P. & Rubin, E. M. (2000). Microarray expression profiling identifies genes with altered expression in hdl-deficient mice. *Genome Research* **10**, 2022–2029.

Cui, X., Gene Hwang, J. T., Qiu, J., Blades, N. J. & Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **6**, 59–75.

Degrelle, S. (2006). Croissance et différenciation du trophoblaste de mammifères en début de gestation: étude par génomique fonctionnelle de l'embryon bovin normal et cloné. PhD thesis, Université de Versailles-Saint-Quentin-en-Yvelines.

Delmar, P., Robin, S., Tronik-Le Roux, D. & Daudin, J.-J. (2005*a*). Mixture model on the variance for the differential analysis of gene expression data. *Applied Statistics* **54**, 31–50.

Delmar, P., Robin, S. & Daudin, J.-J. (2005*b*). VarMixt: efficient variance modelling for the differential analysis of replicated gene expression data. *Bioinformatics* **21**, 502–508.

Feng, S., Wolfinger, R. D., Chu, T. M., Gibson, G. C. & McGraw, L. A. (2006). Empirical Bayes analysis of variance component models for microarray data. *Journal of Agricultural, Biological, and Environmental Statistics* **11**, 197–209.

Foulley, J.-L., San Cristobal, M., Gianola, D. & Im, S. (1992). Marginal likelihood and Bayesian approaches to the analysis of heterogeneous residual variances in mixed linear Gaussian models. *Computational Statistics and Data Analysis* **13**, 291–305.

Gelman, A. (2005). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 1–19.

Kerr, M. K., Martin, M. & Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7**, 819–837.

Kerr, M., Afshari, C., Bennett, L., Bushel, P., Martinez, J., Walker, N. & Churchill, G. (2002). Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica* **12**, 203–217.

Layard, M. W. J. (1973). Robust large-sample tests for homogeneity of variances. *Journal of the American Statistical Association* **68**, 195–198.

Lewin, A., Richardson, S., Marshall, C., Glazier, A. & Aitman, T. (2006). Bayesian modeling of differential gene expression. *Biometrics* **62**, 1–9.

Lönnstedt, I. & Speed, T. (2002). Replicated microarray data. *Statistica Sinica* **12**, 31–46.

Moser, B. K. & Stevens, G. R. (1992). Homogeneity of variance in the two-sample means test. *The American Statistician* **46**, 19–21.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R. & Tsui, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**, 37–52.

Preisser, L., Houot, L., Teillet, L., Kortulewski, T., Morel, A., Tronik-Le Roux, D. & Corman, B. (2004). Gene expression in aging kidney and pituitary. *Biogerontology* **5**, 39–47.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**, 3.

Spiegelhalter, D. J., Thomas, A. & Best, N. G. (2004). *WinBUGS Version 1.4 User Manual*. Cambridge: Medical Research Council Biostatistics Unit. Available from http://www.mrc-bsu.cam.ac.uk/bugs

Tusher, V., Tibshirani, R. & Chu, G. (2001). Significance analysis of microarrays applied to ionizing radiation response. *Proceedings of the National Academy of Sciences of the USA* **98**, 5116–5121.

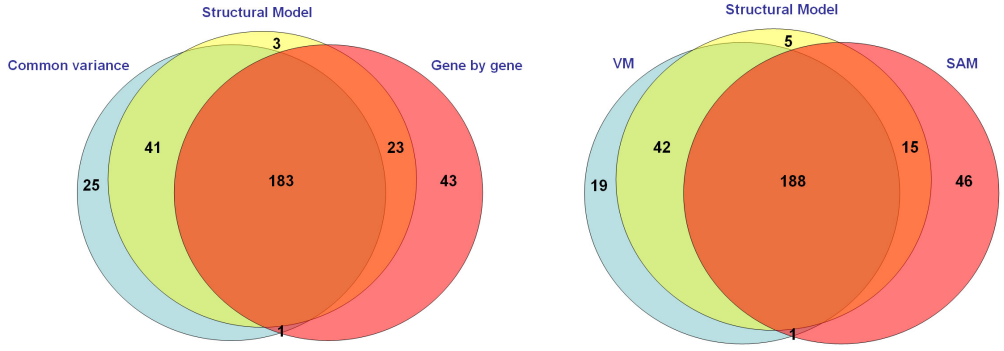
Wright, G. W. & Simon, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* **19**, 2448–2455.

1.2 Complementary results

Most of the results of this paper were obtained at the end of my Master studies at ENSAI (National School of Statistic and Information Analysis) just before the beginning of my PhD. During this internship, I implemented the approximated estimation procedure of the structural model in R and compared it to Varmixt (Delmar *et al.*, 2005) and SAM (Tusher *et al.*, 2001). The analysis with the exact estimation procedure with WinBUGS had already been performed by F. Jafffrézic and J.-.L Foulley. I performed the comparison with the limma model during my PhD after the EADGENE (European Animal Disease Genetics Network of Excellence for Animal Health and Food Safety) workshop on microarray data analysis, where we realised that it was the mostly used package in the microarray community. This EADGENE workshop is presented in de Koning *et al.* (2007) given in Appendix. In brief, the workshop on microarray data analysis was organised in the context of the EADGENE european network. It gathered researchers from 10 countries. All participants received the same two datasets, the first one simulated and the second one coming from real dairy cattle microarray experiments. Among the three other papers written after this workshop, my work contributed to two of them (Watson *et al.*, 2007; Jaffrézic *et al.*, 2007). These papers are also given in Appendix. In this section, only complementary results about the simulated dataset (Watson *et al.*, 2007) are presented since the results for the real dataset have not been biologically validated yet and it is thus more difficult to interpret the results that we found. The simulated dataset offered the advantage to have been simulated independently from people who analysed it, which eliminates the simulation bias which consists in analysing the data with the same model used to generate them. Simulated differentially expressed genes were only known after the workshop. As far as we were concerned, we presented at the workshop results from 6 different models: the structural model SMVar, SAM (Tusher *et al.*, 2001), VM, VM2 (Delmar *et al.*, 2005), the gene-by-gene model and the common variance model. For the paper, we only kept the structural model since it appeared to be a good compromise between the other models. For the workshop in itself, all participants were asked to bring the lists of the 250 top genes detected. That is why my work for this workshop relied on the comparison of such top lists for the 6 models cited before. These models were computed with the siggenes, Varmixt and SMVar packages of the R 2.2.1 version. Between the 6 top lists of differentially expressed genes, 170 genes were found in common. The top list of VM was the closest one to the top list of SMVar with 230 genes in common. The other models had at least 200 genes in common with the structural model : 226 for VM2, 224 for the common variance model, 206

for the gene-by-gene model and 203 for SAM. Figure 1.1 illustrates the fact that SMVar was a good compromise between all models.

Figure 1.1: Venn diagrams for the comparison between the 250 top genes detected with the structural model (SMVar) and two other models (gene-by-gene and common variance model, or VM and SAM).



Only 3 genes that were detected by the structural model were not detected by either the gene-by-gene model or the common variance model, and only 5 genes were not detected by either SAM or VM.

This simulated data set was a very good example to show the performance SMVar had compared to the other models, especially for gene ranking as we only considered the 250 top gene lists. Once the simulated list of differentially expressed genes was given after the workshop, we realised that there were no false positives in any of these top lists. When entire lists of differentially expressed genes at a 5% Benjamini Hochberg threshold were considered, all methods performed quite well (see table 1.1).

Table 1.1: Comparison of lists of differentially expressed genes at 5%-BH - EADGENE simulated data set

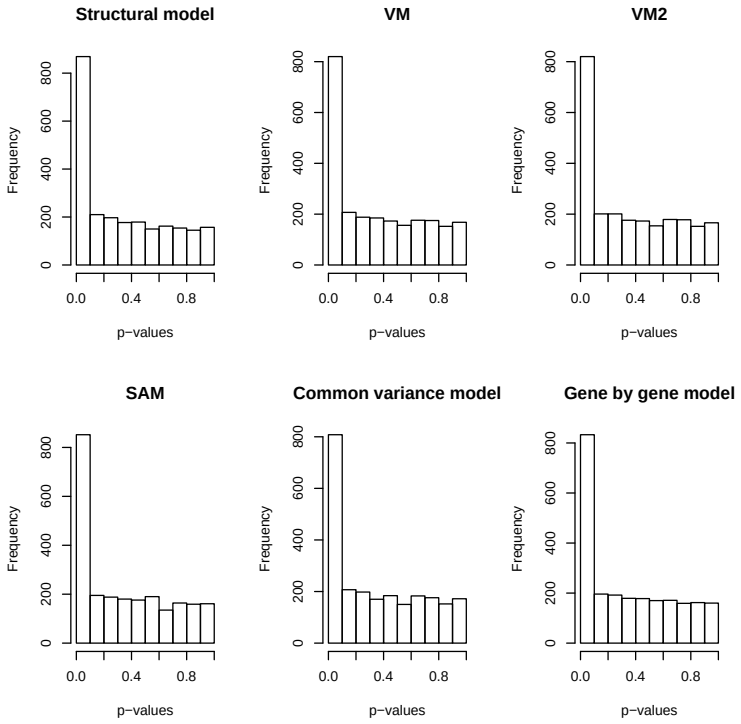
	No	Correct	FP	FN
SMVar	663	614	49	10
VM	646	612	34	12
VM2	648	614	34	10
SAM	660	603	57	21
Common variance	655	615	40	9
Gene-by-gene	626	594	32	30

The best method on this simulated data set was Varmixt. During my

internship, this model was found to perform very well on paired data but can not be easily extended to more than two conditions as the structural model. This is due to the fact that it models the difference between two conditions and does not offer the flexibility to have one variance per condition. SAM was found here to be the worst model as it is the one detecting the most false positives with still a large number of false negatives. Thus, the results on this EADGENE simulated data set confirm the Varmixt and SAM performances on paired data that were already presented in the previous paper (Jaffrézic *et al.*, 2007). The good results obtained here for the common variance and the gene-by-gene models are due to the way data were simulated. In fact, the software SIMAGE (Albers *et al.*, 2006) does not offer the possibility to have heterogeneous variances across genes which considerably advantaged the common variance model. Moreover, ten biological replicates were simulated per condition and because the two technical replicates had a small correlation, they could be considered as independent. Since twenty measures per gene is large, this favoured the gene-by-gene model. One way to show that all models were adequate for this simulated data set is to look at the histograms of raw p-values given in figure 1.2.

If the assumptions made by the model are reasonnable, the p-value distribution has to be uniform for non differentially expressed genes, which is the case for all models here. The pick near 0 represents the p-values of differentially expressed genes. The histogram corresponding to the common variance model is far from the one observed on the real data set of Jaffrézic *et al.* (2007). This illustrates the fact that all models were adapted for these simulations and that the way of simulating data did not seem to be the most appropriate to compare variance modellings. This simulated dataset, however, confirmed the advantages of the structural model, showing that its top differentially expressed gene list was a good compromise between the other models.

Figure 1.2: Histograms of raw p-values on the EADGENE simulated data set

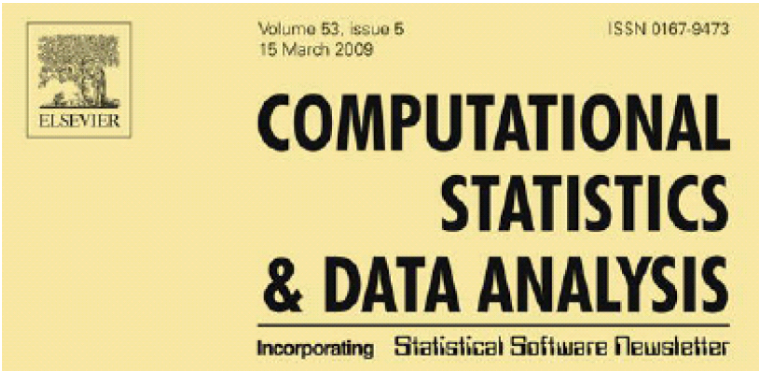


Chapter 2

Modelling of covariance matrices

After the promising results of the structural model on variances, it was natural to extend it to covariance matrices. What is more, biologists from INRA Nouzilly (Benoit Guyonnet, Jean-Luc Gatti) brought us a dataset where repeated measures were performed on the same individuals. They were studying the pig fertility and were looking for differentially expressed genes along the epididymis (long tube where spermatozoa acquire maturation). They cut the epididymis into eleven zones and looked for differentially expressed genes between zones. Samples for different zones were harvested from the same individuals. More details about this project can be found in Guyonnet (2008). From the statistical point of view, it was important to model the covariance matrix between zones. The approaches we proposed are presented in the following paper published in the CSDA special issue 'Statistical Genetics & Statistical Genomics'

2.1 A structural mixed model to shrink covariance matrices



Computational Statistics and Data Analysis 53 (2009) 1630–1638

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A structural mixed model to shrink covariance matrices for time-course differential gene expression studies

Guillemette Marot *, Jean-Louis Foulley, Florence Jaffrézic

INRA, UR 337, F-78350 Jouy-en-Josas, France

ARTICLE INFO

Article history:
Available online 23 April 2008

ABSTRACT

Time-course microarray studies require a particular modelling of covariance matrices when measures are repeated on the same individuals. Taking into account the within-subject correlation in the test statistics for differential gene expression, however, requires a large number of parameters when a gene-specific approach is used, which often results in a lack of power due to the small number of individuals usually considered in microarray experiments. Shrinkage approaches can improve this detection power in differential gene expression studies by reducing the number of parameters, while offering a good flexibility and a small rate of false positives. A natural extension of the shrinkage approach based on a structural mixed model to variance–covariance matrices is proposed. The structural model was used in three configurations to shrink (i) the eigenvalues in an eigenvalue/eigenvector decomposition, (ii) the innovation variances in a Cholesky decomposition, (iii) both the variances and correlation parameters of a gene-by-gene covariance matrix using a Fisher transformation. The proposed methods were applied both to a publicly available data set and to simulated data. They were found to perform well, compared to previously proposed empirical Bayesian approaches, and outperformed the gene-specific or common-covariance methods in many cases.

1. Introduction

Microarray experiments have been widely used over the past few years to study the expression of thousands of genes simultaneously and to detect differentially expressed genes under various conditions. The understanding of many biological processes such as embryonic development or evolution of a disease often relies on time-course experiments. Detection of differentially expressed genes is an essential preliminary step to expression profile clustering and gene network studies in order to reduce the number of genes and focus only on the biologically relevant ones.

Differential expression in time-course experiments has been studied from different points of view. Two main approaches have been considered. First, several authors have focussed on differential gene expression profiles (Storey et al., 2005; Conesa et al., 2006; Angelini et al., 2007). In this framework, they consider global trends, using for example, spline functions, rather than differential gene expression between specific time points. This approach is especially useful when a large number of measurement times has to be analysed.

An alternative to differential expression profiles is to use modified F-tests, which are an extension of the modified t-tests developed for classical differential expression studies (Tusher et al., 2001; Smyth, 2004; Delmar et al., 2005; Jaffrézic et al., 2007). While paired modified t-tests have been proposed in these papers and are well adapted for comparing two correlated

* Corresponding address: INRA, UR 337 Station de Génétique Quantitative et Appliquée, F-78350 Jouy-en-Josas, France. Tel.: +33 1 34 65 21 90; fax: +33 1 34 65 22 10.
E-mail address: guillemette.marot@jouy.inra.fr (G. Marot).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.04.018

conditions, modified F-tests have until now been less investigated. They are, however, the method of choice to analyse time-course experiments, since they allow comparison of three or more conditions with a complex covariance structure.

In order to take into account the within-subject correlation for the analysis of longitudinal gene expression data, Guo et al. (2003) proposed a robust Wald statistic based on an extension of the generalised estimating equations (Liang and Zeger, 1986), using a working correlation matrix. To avoid singularity problems, they suggested adding in the test statistic, similar to the SAM approach (Tusher et al., 2001), a diagonal matrix with positive elements.

A modified F-test was also presented by Smyth (2004) and is implemented in the Bioconductor R package 'Limma'. They proposed including a subject effect in the linear model fitted to the data, and constructing the F-test statistics based on empirical gene-by-gene covariance matrices with shrunk diagonal variance terms. Variance parameters were shrunk, as in the univariate case, using an empirical Bayesian approach.

The aim of this paper was to propose several other F-type statistics for time-course microarray studies based on an extension of the structural mixed model presented by Jaffrézic et al. (2007) to the multivariate case. We focussed on two main decompositions of the empirical gene-by-gene covariance matrices: eigenvalue/eigenvector and Cholesky decomposition. A structural mixed model was used in three configurations to shrink (i) the eigenvalues, (ii) the innovation variances, (iii) both the variances and the correlation coefficients. The F-statistics based on these shrunk covariance matrices were compared to a gene-by-gene analysis, a common covariance model and the modified Limma F-test both on simulated and real data sets.

2. Materials and methods

Let y_{grt} be the expression level for gene g ($g = 1, \ldots, G$), replicate r ($r = 1, \ldots, R_g$) and time t ($t = 1, \ldots, T$). Data are assumed to have been previously normalised.

$$\boldsymbol{Y_{gr}} \sim \mathcal{N}(\boldsymbol{\mu_g}, \boldsymbol{\Sigma_g}), \tag{1}$$

with $\boldsymbol{Y_{gr}} = (y_{gr1}, \ldots, y_{grt}, \ldots, y_{grT})'$ and $\boldsymbol{\mu_g} = (\mu_{g1}, \ldots, \mu_{gt}, \ldots, \mu_{gT})'$. The test hypotheses for differential expression of gene g between several times can be written as H_0: $\boldsymbol{L\mu_g} = 0$ vs H_1: $\boldsymbol{L\mu_g} \neq 0$ with $\boldsymbol{L}$ a given contrast matrix. The Wald statistic can be written as

$$W_g = (\boldsymbol{L}\widehat{\boldsymbol{\mu_g}})'[\boldsymbol{L}\widehat{\boldsymbol{V_g}}\boldsymbol{L}]^{-1}(\boldsymbol{L}\widehat{\boldsymbol{\mu_g}}), \tag{2}$$

where $\widehat{\boldsymbol{\mu_g}}$ and $\widehat{\boldsymbol{V_g}}$ are the maximum likelihood estimation of $\boldsymbol{\mu_g}$ and its sampling variance, respectively.

2.1. Gene-by-gene and common covariance models

A first possibility is to perform a simple gene-by-gene analysis. In this case, $\widehat{\boldsymbol{\mu_g}} = \bar{\boldsymbol{Y}}_{\boldsymbol{g}} = (\bar{y}_{g1}, \ldots, \bar{y}_{gt}, \ldots, \bar{y}_{gT})'$, with $\bar{y_{gt}} = \frac{1}{R_g}\sum_{r=1}^{R_g} y_{grt}$, and matrix $\boldsymbol{V_g}$ is estimated by $\widehat{\boldsymbol{S_g}}/R_g$, where $\widehat{\boldsymbol{S_g}}$ is the empirical variance–covariance matrix $\widehat{\boldsymbol{S_g}} = (R_g - 1)^{-1}\sum_{r=1}^{R_g}(\boldsymbol{Y_{gr}} - \bar{\boldsymbol{Y}}_{\boldsymbol{g}})(\boldsymbol{Y_{gr}} - \bar{\boldsymbol{Y}}_{\boldsymbol{g}})'$.

The test statistic is then defined as (Rao, 1973):

$$F_g^* = \lambda_g(W_g/q), \tag{3}$$

with $\lambda_g = \frac{R_g - q}{R_g - 1}$ and $q = \text{rank}(\boldsymbol{L})$. Under the null hypothesis, $F^* \sim \mathit{Fisher}(q, R_g - q)$. The main difficulty for this approach in microarray studies is the estimation of a large number of parameters, due to the large number of genes analysed simultaneously, with only a few biological replicates. This issue tends to lead to a lack of detection power for the gene-by-gene approach.

On the contrary, a common covariance structure could be assumed for all genes, considering for example the mean of the empirical gene-by-gene covariance matrices. This increases the detection power but also considerably increases the number of false positives when the homogeneity assumption is not fulfilled.

To overcome these drawbacks, various shrinkage methods have been proposed in the literature (Cui et al., 2005; Lewin et al., 2006; Baldi and Long, 2001) when independence was assumed between conditions. We propose here to extend the structural mixed model approach presented by Jaffrézic et al. (2007) to time-course microarray studies to shrink variance–covariance matrices.

2.2. Structural model for both variances and correlation parameters

The empirical gene-by-gene covariance matrix can be written as:

$$\widehat{\boldsymbol{S_g}} = \boldsymbol{D_g}^{1/2}\boldsymbol{R_g}\boldsymbol{D_g}^{1/2}, \tag{4}$$

where $\boldsymbol{D_g}$ is a diagonal matrix of dimension $(T \times T)$ with the empirical variances as diagonal terms and $\boldsymbol{R_g}$ is the empirical gene-by-gene correlation matrix. Let $\ln(\boldsymbol{D_g}) = \text{diag}(\ln(d_{g1}^2), \ldots, \ln(d_{gT}^2))$. The most straightforward extension of the

structural model approach (Jaffrézic et al., 2007; Foulley et al., 1992) to covariance matrices is to shrink the empirical gene-by-gene variances using the following model:

$$\ln d_{gt}^2 = \mu_t + \delta_{gt}, \tag{5}$$

where μ_t is a condition (or time) effect, assumed fixed, and δ_{gt} is a gene effect in condition t. Gene effects are assumed to be independent and normally distributed with mean zero and variance τ_t^2, i.e. $\delta_{gt} \sim \mathcal{N}(0, \tau_t^2)$.

In order to extend the structural approach to shrink the correlation terms we propose to apply a Fisher transformation to gene-by-gene empirical correlations r_{gst} (for times s and t) such as:

$$\rho_{gst} = \frac{1}{2} \ln \left(\frac{1 + r_{gst}}{1 - r_{gst}} \right) = \operatorname{arcth}(r_{gst}). \tag{6}$$

The proposed structural mixed model to shrink the correlation terms is then:

$$\rho_{gst} = \mu_{st} + \delta_{gst}, \tag{7}$$

where $\delta_{gst} \sim \mathcal{N}(0, \tau_{st}^2)$.

2.3. Structural model for an eigenvalue decomposition

Another possible extension of the structural mixed model to covariances would be to shrink the eigenvalues via an eigenvalue/eigenvector decomposition. Let $\widehat{\boldsymbol{S}_g} = \boldsymbol{U_g \Delta_g U_{g'}}$ be the eigenvalue decomposition of the empirical gene-by-gene covariance matrices, where $\boldsymbol{\Delta_g}$ is a diagonal matrix of dimension $(T \times T)$ with eigenvalues λ_{gt} as elements. The columns of matrix $\boldsymbol{U_g}$ correspond to the eigenvectors of $\widehat{\boldsymbol{S_g}}$. A structural mixed model can easily be applied to shrink the eigenvalues across genes such as: $\ln \lambda_{gt} = \mu_t + \delta_{gt}$, where μ_t is a time effect, assumed fixed, and δ_{gt} is a gene effect in time t. As previously, gene effects are assumed independent and normally distributed with mean zero and variance ν_t^2 to be estimated, i.e. $\delta_{gt} \sim \mathcal{N}(0, \nu_t^2)$. This shrinkage of the eigenvalues is also quite similar to the approach presented by Daniels and Kass (2001).

2.4. Structural model for a Cholesky decomposition

A third possibility to shrink the empirical variance–covariance matrices is to apply the Cholesky decomposition as in Daniels and Pourahmadi (2002):

$$\boldsymbol{T_g \widehat{S}_g T_g'} = \boldsymbol{W_g}, \tag{8}$$

where $\boldsymbol{T_g}$ is a lower triangular matrix with 1s on the diagonal and $\boldsymbol{W_g}$ is a diagonal matrix with positive diagonal entries. In the antedependence models (Gabriel, 1962), terms of matrix $\boldsymbol{T_g}$ correspond to antedependence parameters, and terms of matrix $\boldsymbol{W_g}$ are the innovation variances w_{gt}^2, as defined below. The antedependence model for gene g can be written as:

$$y_{grt} = \sum_{j=1}^{t-1} \phi_{gjt}\, y_{grj} + \epsilon_{grt}, \tag{9}$$

where $\epsilon_{grt} \sim \mathcal{N}(0, w_{gt}^2)$. Let $\ln(\boldsymbol{W_g}) = \operatorname{diag}(\ln(w_{g1}^2), \ldots, \ln(w_{gT}^2))$. A structural mixed model can be used for the innovation variances such that for gene g at time t:

$$\ln w_{gt}^2 = \mu_t + \delta_{gt}, \quad \text{with } \delta_{gt} \sim \mathcal{N}(0, \sigma_t^2). \tag{10}$$

For these three configurations of shrinkage, the empirical estimation procedure presented by Jaffrézic et al. (2007) can almost readily be applied. The residual variance in this approximated estimation approach will be fixed to $2/(R_g - 1)$ when working on the log of the eigenvalues or variances, and equal to $1/(R_g - 3)$ for the correlation parameters using the Fisher transformation.

3. Application

The proposed extensions of the structural model for time-course microarray studies were applied to a real data set, publicly available in the GEO (Gene Expression Omnibus) database (http://www.ncbi.nlm.nih.gov/projects/geo/). Its GEO accession number is GSE1440. Human UVC (short-wavelength UV light) irradiated RKO cells were collected and RNA from either whole cell or polysomal fractions was extracted at 0 (control), 3, 6 and 12 h after irradiation. RNA, after reverse transcription and radiolabelling, was hybridised on human cDNA arrays. The experiment was repeated independently three times (A, B, C). We only used here the normalised data from the polysomal fractions of experiment A. We therefore had 11 replicates at each time, each replicate coming from one of the 11 polysomal fractions. There were in total 9600 genes. Since the samples were collected on the same polysomal fractions across time, one can expect a high correlation between time. The different

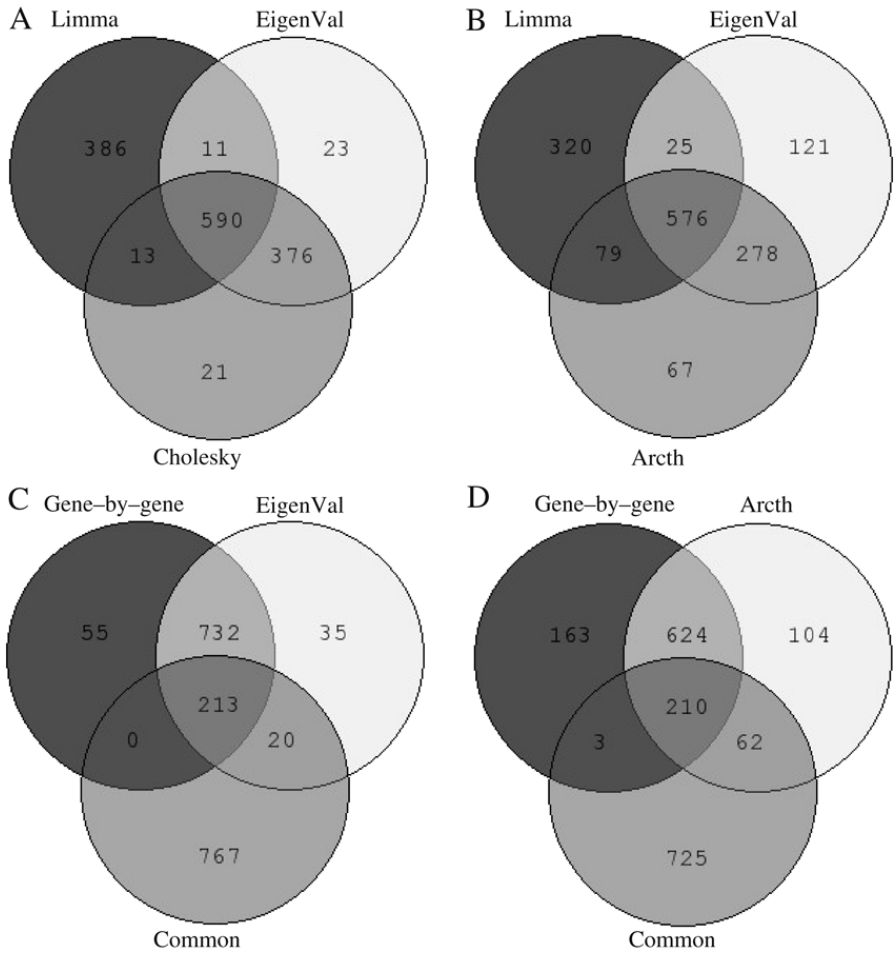


Fig. 1. Venn diagrams comparing the top 1000 gene lists obtained with the different methods on the real data set GSE1440. EigenVal: corresponds to the eigenvalue decomposition of the covariance matrices with a shrinkage of the eigenvalues using a structural mixed model; Cholesky: corresponds to a Cholesky decomposition where the innovation variances were shrunk using a structural model; Arcth: shrinkage of both variances and correlation parameters; Gene-by-gene: corresponds to gene-by-gene empirical covariance matrices; Common: assumes a common covariance structure for all the genes; Limma: modified F-test implemented in the Limma Bioconductor R library (Smyth, 2004) with an individual effect previously fitted in the linear model.

covariance modellings presented above were applied to this data set to find differentially expressed genes between at least two of the four times. They were compared to the gene-by-gene approach, the common covariance model and the modified F-test computed in the Limma R library. The top 1000 gene lists, which correspond to the number of differentially expressed genes found with Limma at a 1% Benjamini–Hochberg threshold, were compared and Venn diagrams are presented in Fig. 1.

It was found here that the common covariance model provided a list for the top 1000 genes that was very different from the other methods. Indeed, only 187 genes were found in common between all six models whereas 554 genes were common between the five other models. The assumption of a common covariance structure for all the genes may therefore not be realistic for this data set. As shown in Fig. 1B, the model based on a Fisher transformation to shrink both the variances and correlations (“Arcth”) appears as a good compromise between Limma and the model based on eigenvalue decomposition (“EigenVal”). In fact, only 67 genes in the top 1000 gene list of Arcth are not detected either by Limma or EigenVal (Fig. 1B). The shrinkage approach based on a Cholesky decomposition was found to provide very similar results to the eigenvalue model since only 44 genes differed between the two methods (Fig. 1A). Both methods were also quite close to the Limma approach with almost 600 genes in common among the top 1000 genes (Fig. 1A). These methods were also found in this example to be quite close to the gene-by-gene model (Fig. 1C), which is due to the quite large number of replicates available in this data set (11 replicates per time).

Since the real top 1000 genes were not known for this data set, it was difficult to assess which modelling was the best. In the next section we therefore present a simulation study which was aimed at evaluating the influence of heterogeneity of the covariances across genes, the number of times and replicates on the performances of the different models.

4. Simulation study

Two large sets of simulations were conducted. The first one compared the different models for various numbers of measurement times (3, 4 and 5 times) and various numbers of replicates per time (6, 8 and 11 replicates). In this first

set of simulations, individual sampling covariance matrices were simulated gene by gene using a Wishart distribution. In the second set of simulations, an additional between-gene heterogeneity of covariances was considered using an inverse Wishart distribution as described below. For each data set, 3000 covariance matrices were simulated corresponding to 3000 genes. Among them, 100 genes were simulated to be differentially expressed between the first time point and another time, with an equal proportion for all time points. For these genes, the logratio between the first time and the other time point was simulated with a uniform [0.5, 2.0].

4.1. Simulation of the covariance matrices

For the first set of simulations, 3000 covariance matrices were simulated with a Wishart distribution of parameters (d, m) where d is the number of degrees of freedom usually equal to $(R-1)$, with R the number of replicates. Depending on the number of times considered, matrix m equals m_3, m_4 or m_5, as defined below. Each simulated matrix was then divided by d to be close to the original m matrix.

$$m_3 = \begin{pmatrix} 1 & 0.8 & 0.6 \\ 0.8 & 1 & 0.8 \\ 0.6 & 0.8 & 1 \end{pmatrix} \qquad m_4 = \begin{pmatrix} 1 & 0.8 & 0.6 & 0.5 \\ 0.8 & 1 & 0.8 & 0.6 \\ 0.6 & 0.8 & 1 & 0.8 \\ 0.5 & 0.6 & 0.8 & 1 \end{pmatrix} \qquad m_5 = \begin{pmatrix} 1 & 0.8 & 0.6 & 0.5 & 0.4 \\ 0.8 & 1 & 0.8 & 0.6 & 0.5 \\ 0.6 & 0.8 & 1 & 0.8 & 0.6 \\ 0.5 & 0.6 & 0.8 & 1 & 0.8 \\ 0.4 & 0.5 & 0.6 & 0.8 & 1 \end{pmatrix}.$$

In the second set of simulations, the variability between genes was simulated with an inverse Wishart distribution of parameters ν and m. In order to keep the ratio ρ of between-gene and sampling variability close to a ratio found in a previously analysed real data set (bovine data presented in Jaffrézic et al. (2007)), we chose $\rho = 2/3$. As explained in the Appendix, parameter ν was then defined as $\nu = \max(d/2, (T+2))$. To obtain simulated matrices centered around m, all the inverse Wishart matrices were multiplied by $(\nu - T - 1)$. Then, one observation was simulated for each of these 3000 new matrices $(M_1, \ldots, M_{3000})$ with a Wishart distribution of parameter $(d, M_i)_{(i=1,\ldots,3000)}$ in order to simulate an additional sampling variability. Once again, each matrix obtained was divided by d to be close to the original m structure.

4.2. Results

As Opgen-Rhein and Strimmer (2007), we used receiver-operator characteristic (ROC) curves to compare the ranking of the genes obtained with the different methods. For this, we computed the number of False Positives (FP), True Positives (TP), False Negatives (FN) and True Negatives (TN) for all possible cut-offs in the gene list (1-200). This procedure was repeated 150 times for each test statistic to obtain estimates of the ROC curves describing the dependency between sensitivity $E(\frac{TP}{TP+FN})$ and specificity $E(\frac{TN}{TN+FP})$. The results from the first set of simulations are given in Fig. 2. The best model is the one which maximises the area under the curve.

Since only sampling variability was simulated in this first set of data using a standard Wishart distribution, the model that was found to fit these data the best was the common covariance model for any number of times and replicates considered here. Among the proposed shrinkage approaches, the model based on the Fisher transformation where both the variances and correlation parameters were shrunk using a structural model ("Arcth") appeared to perform the best in this case. Limma was found to perform quite well for small numbers of replicates. For more than 8 replicates, however, it did not seem to be quite appropriate. In fact, even the gene-by-gene model was found to perform better for more than 8 replicates at 4 or 5 times. These simulations confirmed, as observed on the real data set, that the proposed methods based either on the shrinkage of the eigenvalues or innovation variances provide quite similar results. Both were found to perform better here than the simple gene-by-gene approach and were also found to perform better than Limma for 8 replicates and more.

ROC curves corresponding to the second set of simulations are presented in Fig. 3. It can be observed that between-gene heterogeneity induced by the inverse Wishart distribution has a direct consequence on the performance of the common covariance model, meaning that it now becomes the worst model whatever the number of times and replicates considered. Limma also shows quite poor results in this situation of large covariance heterogeneity. The shrinkage approach where both the variances and correlation parameters were shrunk, although better than Limma for most cases, was not found to perform better than the gene-by-gene approach. On the other hand, the proposed methods based either on the shrinkage of the eigenvalues or of the innovation variances were found to perform better than all the other methods for any number of times or replicates considered.

5. Discussion

The extension of the structural mixed model on variances to variance–covariance matrices was a natural way to take into account the within-subject correlations in time-course microarray studies. The structural model proved to be particularly useful when the number of times and replicates increased (more than 4 times, more than 8 replicates). The two sets of simulations showed that the ranking of methods was highly dependent on the degree of heterogeneity between the covariances across genes. In fact, the simple common covariance model performed well when only sampling variability

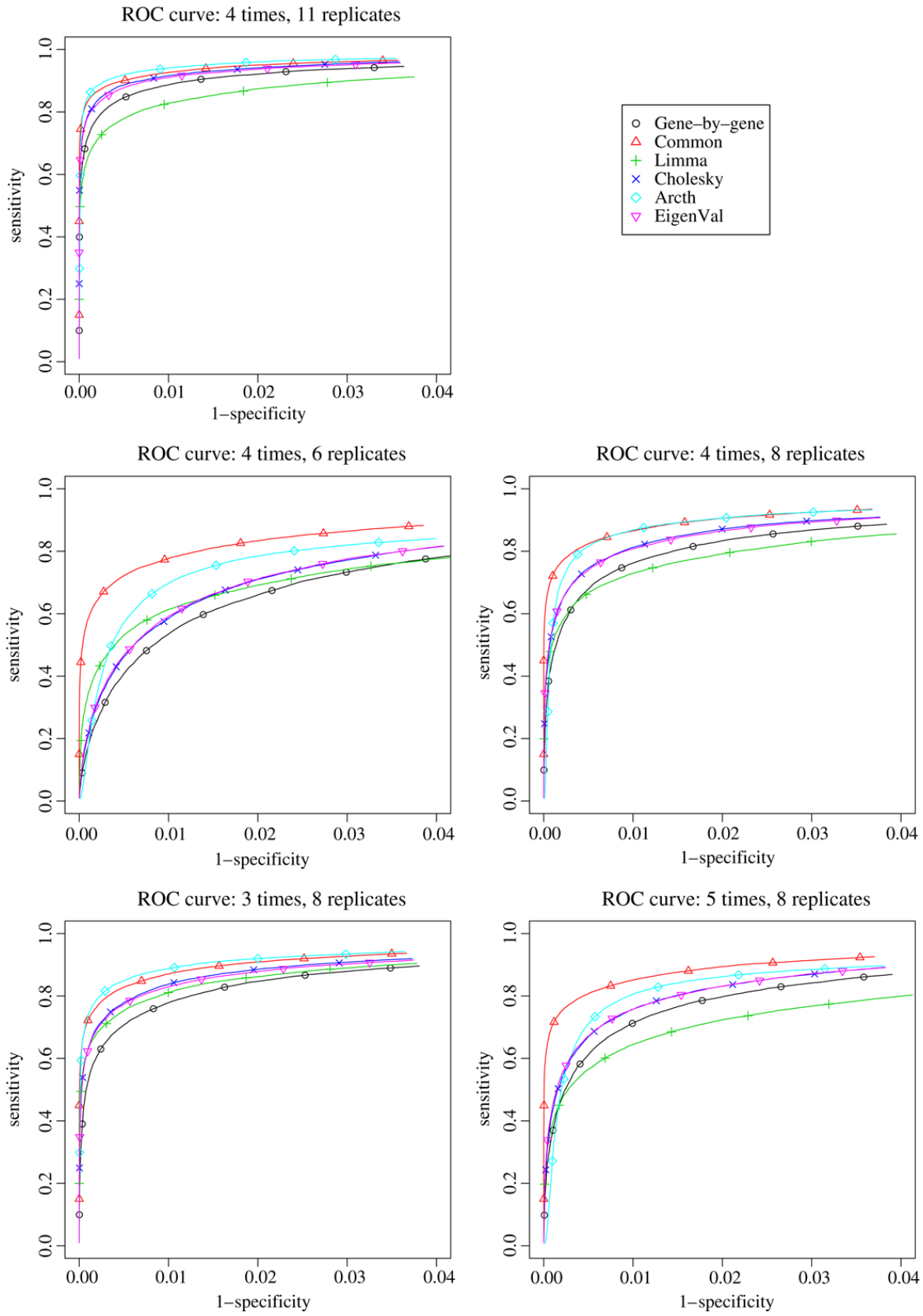


Fig. 2. ROC curves for the first set of simulations with covariance matrices simulated with a Wishart distribution. Names of the different methods are the same as in Fig. 1.

was simulated but very poorly for a large between-gene variability. Similarly, the shrinkage approach on both variance and correlation parameters behaved better when little heterogeneity was simulated, but still outperformed Limma in the case of

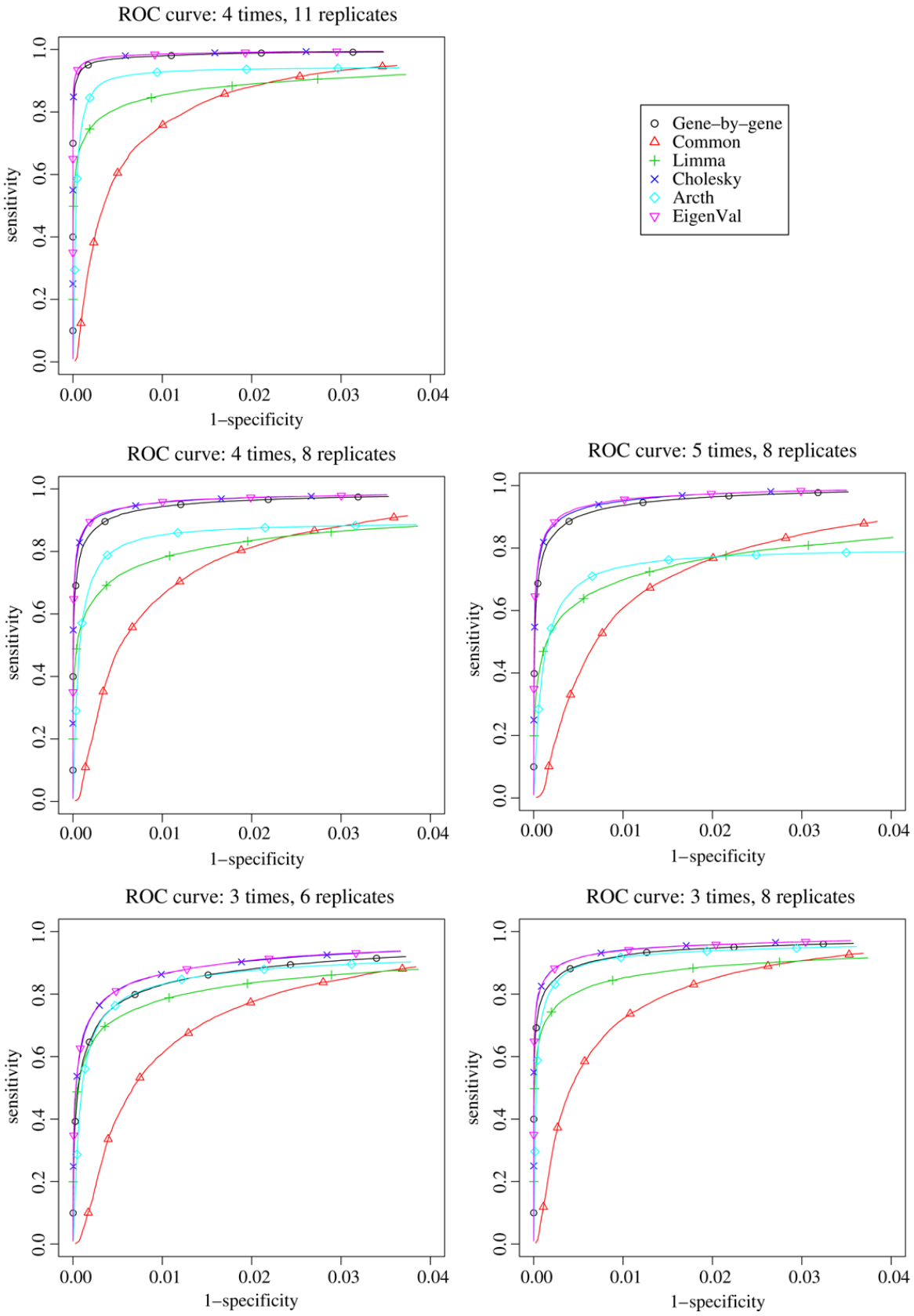


Fig. 3. ROC curves for the second set of simulations where between-gene heterogeneity of covariances was simulated with an additional inverse Wishart distribution. Names of the different methods are the same as in Fig. 1.

higher covariance heterogeneity. The proposed methods based on the shrinkage of the eigenvalues or innovation variances were found, in this simulation study, to be able to adapt to a higher degree of between-gene covariance heterogeneity. The

advantage of the Cholesky decomposition over the eigenvalue decomposition can be seen for ordered time points since the parameters can then have a biological interpretation in terms of antedependence coefficients (Daniels and Pourahmadi, 2002). It can also be noted that the proposed shrinkage approach could be extended to structured antedependence models (Nunez-Anton and Zimmerman, 2000), which would allow to further reduce the number of parameters to estimate. As proposed by Daniels and Kass (2001), it could also be possible to extend the shrinkage approach in the eigenvalue decomposition to both eigenvalues and eigenvectors, which is expected to improve the modelling ability for a smaller degree of heterogeneity. The extension of the structural approach in this case is, however, not straightforward and remains an area of investigation.

The estimation procedure used here for the shrinkage parameters was an empirical Bayesian approach as presented by Jaffrézic et al. (2007). It is based on the calculation of empirical variance–covariance matrices, which can sometimes be close to singularity as pointed out by Guo et al. (2003) due to the small number of replicates in microarray experiments. Furthermore, an important issue in longitudinal data analysis is the problem of missing values. In the context of microarray experiments it is expected, however, that most of the missing values will be due to technical problems and will therefore be considered as "missing completely at random". In this case, the proposed empirical approach will still be consistent. To overcome the problem of singularity or in the case of too large a number of missing values, a fully Bayesian estimation procedure could be used based on MCMC methods, as implemented for variances by Jaffrézic et al. (2007) with the Winbugs software (Spiegelhalter et al., 2004), which will provide more robust estimates. These estimation procedures are, however, much more time consuming.

Another issue that needs to be further investigated is the calculation of the degrees of freedom of the Fisher distribution under the null hypothesis when using these structural shrunk covariance matrices. In fact, although Smyth (2004) was able to directly refer to a number of degrees of freedom calculated for the modified t-test statistic in the modified F-test, we found it very difficult to extend the usual approximations (Satterthwaite's approach used by Jaffrézic et al. (2007), or Kenward and Roger (1997)'s procedure) to our shrinkage methods in the multivariate case. In order to obtain the p-values based on the proposed modified F-statistics with the structural models we would therefore advise to use permutations. These were, however, too time consuming for the extensive simulation study presented here, which is the reason why we based the comparison of methods on ROC curves.

It can be further noted that the proposed covariance modellings could also be useful in differential expression profile studies where very simple covariance structures are usually considered (either diagonal or with only a simple random intercept).

Appendix

A.1. Choice of the parameter of the inverse Wishart distribution

To choose an appropriate value for parameter ν, the aim is to keep ratio ρ of the between-gene and sampling variability close to a ratio found in a real data set. The link between parameters ρ and ν can be calculated as:

$$\rho = \frac{\sigma_g^2}{\sigma_g^2 + 2/d}, \tag{11}$$

where σ_g^2 is the between-gene variability and d is usually equal to $(R-1)$ where R is the number of biological replicates. This is equivalent to writing:

$$\sigma_g^2 = \frac{2\rho}{d(1-\rho)}. \tag{12}$$

On the other hand, parameter ν of the Wishart distribution can be approximated by $2/\sigma_g^2$ (Foulley et al., 2004), i.e.

$$\nu = \frac{1-\rho}{\rho} d. \tag{13}$$

Since we chose $\rho = 2/3$ (as in a microarray experiment analysed before for the detection of differentially expressed genes in bovine embryos (Jaffrézic et al., 2007)), then $\nu = d/2$ and in the simulations, we used

$$\nu = \max(d/2, (T+2)),$$

where T is the total number of times.

References

Angelini, C., De Canditiis, D., Mutarelli, M., Pensky, M., 2007. A bayesian approach to estimation and testing in time-course microarray experiments. Statistical Applications in Genetics and Molecular Biology 6 (1), 24.

Baldi, P., Long, A., 2001. A bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. Bioinformatics 17, 509–519.

Conesa, A., Nueda, M.J., Talon, M., 2006. Masigpro: A method to identify significantly differential expression profiles in time-course microarray experiments. Bioinformatics 22 (9), 1096–1102.
Cui, X., Gene Hwang, J., Qiu, J., Blades, N., Churchill, G., 2005. Improved statistical tests for differential gene expression by shrinking variance components estimates. Biostatistics 6, 59–75.
Daniels, M.J., Kass, R.E., 2001. Shrinkage estimators for covariance matrices. Biometrics 57 (4), 1173–1184.
Daniels, M.J., Pourahmadi, M., 2002. Bayesian analysis of covariance matrices and dynamic models for longitudinal data. Biometrika 89 (3), 553–566.
Delmar, P., Robin, S., Daudin, J.J., 2005. Varmixt: Efficient variance modelling for the differential analysis of replicated gene expression data. Bioinformatics 21 (4), 502–508.
Foulley, J.-L., San Cristobal, M., Gianola, D., Im, S., 1992. Marginal likelihood and bayesian approaches to the analysis of heterogeneous residual variances in mixed linear gaussian models. Computational Statistics and Data Analysis 13, 291–305.
Foulley, J.-L., Sorensen, D., Robert-Granié, C., Bonaiti, B., 2004. Heteroskedasticity and structural models for variances. Journal of Indian Society of Agricultural Statistics 57, 64–70.
Gabriel, K.R., 1962. Ante-dependence analysis of an ordered set of variables. The Annals of Mathematical Statistics 33 (1), 201–212.
Guo, X., Qi, H., Verfaillie, C.M., Pan, W., 2003. Statistical significance analysis of longitudinal gene expression data. Bioinformatics 19 (13), 1628–1635.
Jaffrézic, F., Marot, G., Degrelle, S., Hue, I., Foulley, J.-L., 2007. A structural mixed model for variances in differential gene expression studies. Genetical Research 89 (1), 19–25.
Kenward, M.G., Roger, J.H., 1997. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 53, 983–997.
Lewin, A., Richardson, S., Marshall, C., Glazier, A., Aitman, T., 2006. Bayesian modeling of differential gene expression. Biometrics 62, 1–9.
Liang, K., Zeger, S., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.
Nunez-Anton, V., Zimmerman, D.L., 2000. Modeling nonstationary longitudinal data. Biometrics 56 (3), 699–705.
Opgen-Rhein, R., Strimmer, K., 2007. Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. Statistical Applications in Genetics and Molecular Biology 6 (1), 9.
Rao, C., 1973. Linear Statistical Inference and its Applications, 2nd edition. Wiley, New York.
Smyth, G.K., 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 3 (1), 3.
Spiegelhalter, D., Thomas, A., Best, N., 2004. WinBUGS Version 1.4 User Manual. Cambridge: Medical Research Council Biostatistics Unit. Available from http://www.mrc-bsu.cam.ac.uk/bugs.
Storey, J.D., Xiao, W., Leek, J.T., Tompkins, R.G., Davis, R.W., 2005. Significance analysis of time course microarray experiments. Proceedings of the National Academy of Sciences of the United States of America 102 (36), 12837–12842.
Tusher, V.G., Tibshirani, R., Chu, G., 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America 98 (9), 5116–5121.

2.2 Complementary results

This paper essentially treated the gene ranking without giving any p-values resulting from the model proposed. These p-values, after adjustement for multiple testing, would have been interesting to evaluate the false discovery rate for a given list of genes. However, it was too difficult to calculate the number of degrees of freedom of the distribution under the null hypothesis. Preliminary work about that question is given in 2.2.1. The following subsection deals with the possibility to extend the shrinkage approach to antedependence parameters.

2.2.1 Calculation of the number of degrees of freedom

How to compute Satterthwaite degrees of freedom?

Our first initiative was to make an analogy between the formula presented in equation (7) in Jaffrézic *et al.* (2007) and the one used in SAS (SAS help, version 9.2). Indeed, we thought that compared to our formula, their computing way could be easier to extend from a gene-by-gene approach to a shrinkage approach in the case of the F-test.

We first considered the t-statistic case. The model is:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{Y}$ denotes the vector of observed values, $\boldsymbol{X}$ is the known fixed effects design matrix, and $\boldsymbol{\beta}$ is the unknown fixed effects parameter vector. While assuming that $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V})$, suppose $\boldsymbol{\theta}$ is the vector of unknown parameters in $\boldsymbol{V}$ and suppose $\boldsymbol{C} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-}$, where $^{-}$ denotes a generalized inverse. Let $\widehat{\boldsymbol{C}}$ and $\widehat{\boldsymbol{\theta}}$ be the corresponding estimates. Consider $\boldsymbol{l}$ a vector defining an estimable linear combination of $\boldsymbol{\beta}$. The Satterthwaite degrees of freedom for the t-statistic

$$t = \frac{\boldsymbol{l}\widehat{\boldsymbol{\beta}}}{\sqrt{\boldsymbol{l}\widehat{\boldsymbol{C}}\boldsymbol{l}'}}$$

is computed in SAS as

$$\nu = \frac{2(\boldsymbol{l}\widehat{\boldsymbol{C}}\boldsymbol{l}')^2}{\boldsymbol{g}'\boldsymbol{A}\boldsymbol{g}}$$

where $\boldsymbol{g}$ is the gradient of $\boldsymbol{l}\boldsymbol{C}\boldsymbol{l}'$ with respect to $\boldsymbol{\theta}$, evaluated at $\widehat{\boldsymbol{\theta}}$, and $\boldsymbol{A}$ is the asymptotic variance-covariance matrix of $\widehat{\boldsymbol{\theta}}$ obtained from the second derivative matrix of the likelihood equations.

For the structural model for variances (Jaffrezic et al., 2007), we computed for each gene the number of degrees of freedom ν_g as:

$$\nu_g = \frac{2(\boldsymbol{l}\widehat{\boldsymbol{C}}\boldsymbol{l}')^2}{V(\boldsymbol{l}\widehat{\boldsymbol{C}}\boldsymbol{l}')} \tag{2.1}$$

$$\nu_g = \frac{2(\widehat{\sigma_{gi}}^2 + \widehat{\sigma_{gj}}^2)^2}{V(\widehat{\sigma_{gi}}^2) + V(\widehat{\sigma_{gj}}^2)} \tag{2.2}$$

where $\widehat{\sigma_{gi}}^2$ and $\widehat{\sigma_{gj}}^2$ are the estimators of the variances for gene g in conditions $c = i, j$ and their variances can be approximated by:

$$V(\widehat{\sigma_{gc}}^2) = (\widehat{\sigma_{gc}}^2)^2 V(\widehat{\ln \sigma_{gc}^2}) \tag{2.3}$$

with $V(\widehat{\ln \sigma_{gc}^2}) \approx (1/\tau_c^2 + (R_c - 1)/2)^{-1}$, τ_c^2 the variance of the random gene effect in the structural model and R_c the number of replicates in condition c.

As far as the F-statistic

$$F = \frac{\widehat{\boldsymbol{\beta}}'\boldsymbol{L}'(\boldsymbol{L}\widehat{\boldsymbol{C}}\boldsymbol{L}')^{-1}\boldsymbol{L}\widehat{\boldsymbol{\beta}}}{q}$$

is concerned (with q the rank of the contrast matrix $\boldsymbol{L}$) , it is computed in SAS by first performing the spectral decomposition

$$\boldsymbol{L}\widehat{\boldsymbol{C}}\boldsymbol{L}' = \boldsymbol{P}'\boldsymbol{D}\boldsymbol{P}$$

where $\boldsymbol{P}$ is an orthogonal matrix of eigenvectors and $\boldsymbol{D}$ is a diagonal matrix of eigenvalues, both of dimension $q * q$. Define $\boldsymbol{l_m}$ to be the mth row of $\boldsymbol{PL}$, and let

$$\nu_m = \frac{2(D_m)^2}{\boldsymbol{g_m}'\boldsymbol{A}\boldsymbol{g_m}}$$

where D_m is the mth diagonal element of $\boldsymbol{D}$ and $\boldsymbol{g_m}$ is the gradient of $\boldsymbol{l_m}\boldsymbol{C}\boldsymbol{l_m}'$ with respect to $\boldsymbol{\theta}$, evaluated at $\widehat{\boldsymbol{\theta}}$.

Then, letting

$$E = \sum_m \frac{\nu_m}{\nu_m - 2} I(\nu_m > 2)$$

the degrees of freedom for F are computed as

$$\nu = \frac{2E}{E - q}$$

provided $E > q$; otherwise ν is set to zero.

To be able to know the number of degrees of freedom of any F-statistic, we wanted to calculate the denominator in the F-statistic case with the same idea as for the t-statistic, that is to say to estimate $Var(\boldsymbol{l_m}\widehat{\boldsymbol{C}}\boldsymbol{l_m}')$. Let $f(u) = \boldsymbol{lCl}'$. The idea was to calculate the sampling variability of the estimator of $f(u)$. Considering the following equation,

$$Var(f(u)) = E_Y(Var(f(u)|Y)) + Var_Y(E(f(u)|Y)) \tag{2.4}$$

$$(1) = (2) + (3)$$

it might be easier to calculate (2) by (1) − (3). We tried this approach on data studying the bovine embryo development (Degrelle, 2006) obtained with similar chips and similar pre-processing as in Jaffrézic *et al.* (2007). Four time points were considered. Since embryos were killed when harvesting the RNA, measures at different times were not performed on the same individuals, we thus did not expect a high correlation. That is why this dataset was not used for illustration of our approach in the CSDA paper while it had been investigated for research work before. For the t-statistic case, we only looked at the difference between time 1 and time 4.

To calculate (3), let $v_g = \widehat{f(u)}$. If we assume that variability between replicates for a gene equals to variability between genes within the sample then

$$(3) = \sum_g \frac{(v_g - v)^2}{G - 1}$$

With all v_g calculated on the real data set as the sums of the shrunk variances for time 1 and time 4, I obtained (3)=0.051926.

To calculate (1), I made simulations of $f(u)$ and then calculated its variance. We assumed that $\ln \sigma^2_{gc} \sim \mathcal{N}(\mu_c, \tau^2_c)$

On the bovine embryos data set, $\mu_1 = -1.870210$, $\mu_4 = -2.03337$, $\tau^2_1 = 0.8317$, $\tau^2_4 = 0.5948$. With 10 000 000 simulations, (2)= 0.044.

This number was to be compared to the denominator of the number of degrees of freedom obtained by the SMVar use which equalled 0.0314. When we used the 0.044 denominator we had about 8 degrees of freedom and when we used the 0.0314 denominator we had about 12 degrees of freedom. This difference being too large to validate the approach by simulations, we did not try to apply it to the F-statistic case. The number of degrees of freedom for shrinkage statistics remains an open question.

Satterthwaite's formula for unequal sample sizes

When searching to extend the Satterthwaite approach from the t-statistic case to the F-statistic case, I realised that the formula given in Jaffrézic *et al.* (2007) was a particular case and was valid only for data sets where sample sizes were equal between conditions.

In a more general case, we would have

$$\nu_g = \frac{2(\frac{\widehat{\sigma_{gi}}^2}{n_i} + \frac{\widehat{\sigma_{gj}}^2}{n_j})^2}{\frac{1}{n_i^2}(\widehat{\sigma_{gi}}^2)^2 V(ln\widehat{\sigma_{gi}}^2) + \frac{1}{n_j^2}(\widehat{\sigma_{gj}}^2)^2 V(ln\widehat{\sigma_{gj}}^2)} \tag{2.5}$$

This derives from the following equation:

$$\nu_g = \frac{(\frac{\widehat{\sigma_{gi}}^2}{n_i} + \frac{\widehat{\sigma_{gj}}^2}{n_j})^2}{\frac{\widehat{\sigma_{gi}}^4}{n_i^2}\frac{1}{\nu_i} + \frac{\widehat{\sigma_{gj}}^4}{n_j^2}\frac{1}{\nu_j}} \tag{2.6}$$

where ν_i are the degrees of freedom associated with the calculation of $\widehat{\sigma_{gi}}^2$ Since the model was on the log of the variances,

$$\nu_i = \frac{2}{V(ln\widehat{\sigma_{gi}}^2)}$$

$$\nu_i = \frac{2}{\tau_k^2} + d_i k = \frac{2}{\tau_k^2} + (n-1)$$

The correction given in equation 2.5 has been taken into account since the version 1.2 of the SMVar package.

2.2.2 Shrinkage of antedependence parameters

Another possible extension, not discussed in the paper, is the modelling of the antedependence parameters in the Cholesky decomposition which would seem natural keeping the same idea of shrinking more than only diagonal values. We actually did implement an approximated estimation to shrink antedependence parameters. We modelled the antedependence parameters similarly to the innovation variances that is to say that for gene g, order of antedependence j and time t:

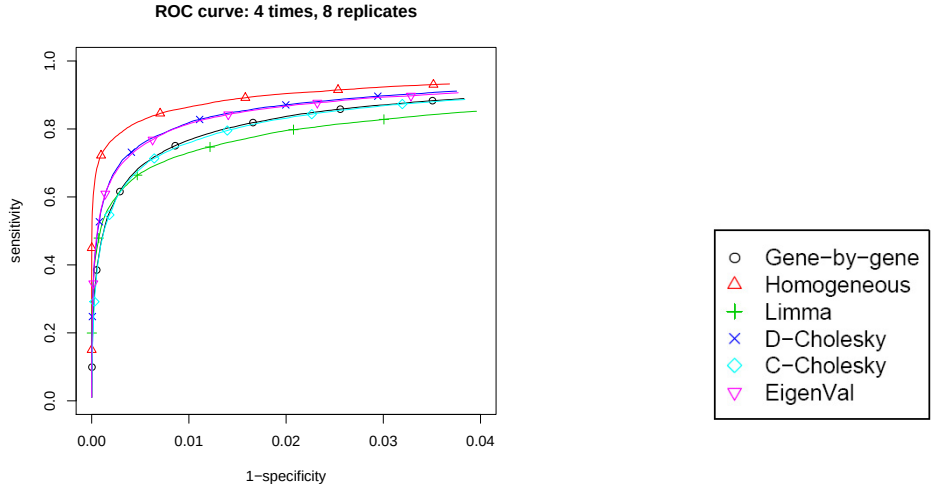
$$\phi_{gjt} = \phi_{jt} + \gamma_{gjt}, \text{ with } \gamma_{gjt} \sim \mathcal{N}(0, \omega_{jt}^2). \tag{2.7}$$

The implementation used to estimate these antedependence parameters was similar to the one used for the logarithms of the innovation variances.

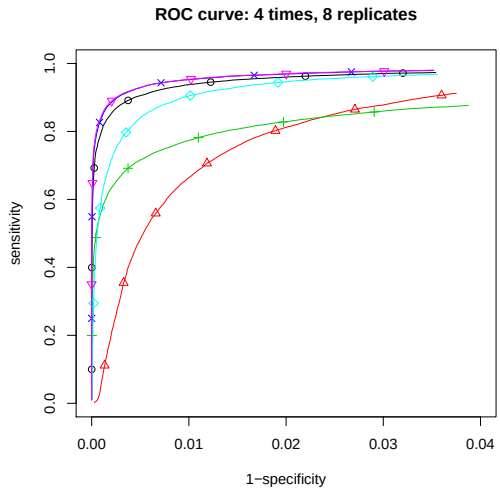
We compared on 150 simulations the ROC curve obtained with the approach shrinking both innovation variances and antedependence parameters with ROC curves from all the models presented in the paper but the one using Fisher transformation. Results are presented in Figure 2.1.

While the results from the set of simulations with heterogeneous covariances are not surprising, that is to say that the C-Cholesky curve stands between the gene-by-gene curve and the common covariance curve, the top graph of ROC curves indicates that there is a problem in the calculation of the shrinkage estimates for the antedependence parameters. Indeed, one would expect that the area under the curve for the approach shrinking both innovation variances and antedependence parameters is higher than the areas under the curves for approaches shrinking only diagonal values since the best model in the case of homogeneity of covariance matrices is the model assuming a common covariance matrix. Actually, we found out that the problem came from the estimation of the residual variance in the structural mixed model assumed on the antedependence parameters. While the residual variance can be fixed to $2/(R_g - 1)$ with R_g the number of replicates when shrinking diagonal values, this estimation is wrong for the shrinkage of antedependence parameters. Since we do not know how to fix the residual variance in this last case, we left out this model and considered the alternative model using Fisher transformation for correlation parameters presented in the paper. In this case, the residual variance is fixed to $1/(R_g - 3)$. To conclude, this example warned us against the abusive use of the structural mixed model with fixing the residual variance. Particular care has to be taken when estimating it.

Figure 2.1: Supplementary ROC curves to compare the shrinkage approach on both the antedependance parameters and innovation variances with the previous models



No heterogeneity of covariances



Heterogeneity simulated with an additional inverse Wishart

EigenVal: corresponds to the eigenvalue decomposition of the covariance matrices with a shrinkage of the eigenvalues using a structural mixed model; D-Cholesky: corresponds to a Cholesky decomposition where the innovation variances (Diagonal terms) were shrunk using a structural model; C-Cholesky: shrinkage of both the innovation variances and antedependence parameters (Complete shrinkage); Gene-by-gene: corresponds to gene-by-gene empirical covariance matrices; Homogeneous: assumes a common covariance structure for all the genes; Limma: modified F-test implemented in the Limma Bioconductor R library (Smyth (2004)) with an individual effect previously fitted in the linear model.

2.3 Application to INRA datasets

Variance-covariance modelling has provided an interesting extension of the structural model for variances presented in Chapter 1. Unfortunately for us, we were not able to apply covariance modelling on the real dataset produced by Benoit Guyonnet and Jean-Luc Gatti (INRA Nouzilly) which initiated the corresponding statistical question. Indeed, since the number of replicates (4 per zone) was lower than the number of measures (11) alongside the epididymis, it was impossible to calculate the gene-by-gene empirical variance-covariance matrices using all measures separately. At first, we gathered several zones together to reduce the total number of zones. This solved the numerical problem of estimating gene-by-gene covariance matrices. Nevertheless, the problem of choosing a cut-off to declare which genes were differentially expressed remained since we did not have any numbers of degrees of freedom for the null distribution. That is why we decided to apply the Limma procedure including an animal effect when looking for differentially expressed genes between at least one zone and the other ones in the pig epididymis real dataset. Concerning the structural model for variances, it was widely applied in this biological study when performing t-tests to find differentially expressed genes between two zones. During my PhD, I built a package SMVar implementing the structural model for variances and shared it with the whole scientific community via the CRAN website. I received some advice from people who used it, which helped me to correct some initial mistakes. For example, my initial version could not handle too many missing data implying a null variance for a given gene. The following versions took into account this possibility excluding genes with a null variance from the analysis. My package was also used by people from INRA Jouy-en-Josas who work on bovine embryos reproduction (Isabelle Hue, Séverine Degrelle, Damien Valour, etc.). This collaboration has resulted in several coauthorships for presentations.

To conclude this part, empirical Bayesian approaches proved to be useful to overcome the problem of small number of individuals in microarray experiments. The following part will consider the high dimensionality problem from another point of view. The small number of samples available is most of the time due to money constraints. In the next part, we will study the possibility to save samples from some experiments to use them for other ones. This takes place in the more general context of sequential analysis.

Part II

Sequential analysis

[0]This part results from my work with C.-D. Mayer (BioSS). Most results were obtained during a 6-month stay in Scotland during my PhD and further collaboration

Chapter 3

Sequential analysis

Sequential analysis refers to statistical analysis where sample size is not fixed in advance. Data are evaluated as they are collected and the experiment is stopped when enough significant results are obtained according to a stopping rule. For those who are not familiar with sequential analysis (especially concerning the stopping of the experiment), details will be given in the first section. This section will also present what sequential analysis usually means, listing and explaining some frequent related words. Then, preliminary results will be shown, extending classical sequential analysis to the field of microarray analysis. These results are not included in the SAGMB (*Statistical Applications in Genetics and Molecular Biology*) paper which follows, since we discovered that sequential analysis was very particular and not as complicated as in the classical case for the microarray data context due to the high dimensionality of these experiments. We however decided to keep some of these results in this PhD report since 1) it explains our research in a chronological way and helps us to better point out why the first results in our paper are so important, 2) it can be useful for other situations where Student statistics are involved without high dimensionality. Section 3.3 consists of the paper accepted in SAGMB and the next one gives complementary results essentially obtained after reviewers' comments.

3.1 Introduction

3.1.1 Terminology of sequential analysis

Sequential analysis was first introduced by Wald as a tool for more efficient industrial quality control during World War II. The principle of the Sequential Probablility Ratio Test (Wald, 1947) and of the triangular test

(Anderson, 1960) is to determine, after each inclusion of a new individual whether the data available are sufficient to choose between H_0 and H_1, two single hypotheses. These first methods were adapted for a comparison of an observed proportion with a theoretical value, and of two proportions. For the last case, the analysis is performed after the inclusion of each pair of subjects. If the collected data are sufficient, no more individuals are added in the analysis. Sequential analysis is then defined as an analysis where data is evaluated as it is collected and further sampling is stopped in accordance with a pre-defined stopping rule as soon as significant results are observed. The stopping rule is the mechanism which tells the statistician when to stop sampling. Sequential testing appeared in the field of medical statistics with the book 'Sequential Medical Trials' (Armitage, 1960). Ideas from this book are given in the overview of Todd (2007). Armitage argued that ethical considerations demand a trial to be stopped as soon as there is clear evidence that one of the treatments is to be preferred. He described a number of techniques and their application to trials comparing two alternative treatments. Ethics is nowadays not the only motivation for sequential analysis. The cost of continuing the experiment is another good reason as sequential analysis often reduces the number of samples needeed. In microarray analysis, the cost of an experiment is the main reason which motivated us to develop a sequential approach for this type of data. After the book of Armitage, clinical trials became a big field of research for sequential analysis. As performing one analysis after each inclusion was not reasonnable, people developed what is now called 'group sequential trials' or 'repeated significance tests'. Designs allowed groups of patients being included between each analysis.

As pointed out in Spiessens *et al.* (2000), two approaches are generally distinguished in group sequential methods. The first one is the 'boundaries' approach of Whitehead and Stratton (1983), where data are monitored continuously. He suggests to monitor trials in terms of information and not sample size. This approach relies on a graphical rule and is based on two statistics 'Z' and 'V'. 'Z' is a measure for the difference between the two treatments and is represented on the vertical axis. 'V' is a measure for the amount of information gathered up to the performed interim analysis and is on the horizontal axis. Graphs and more details are given in the overview of Sébille and Bellissant (2003). The second approach of group sequential methods is the one we considered at first for our microarray analyses. It is based on the repeated significance testing principle where at each interim analysis the significance level is adjusted to control for the overall probability of a type I error (Spiessens *et al.*, 2000). Armitage *et al.* (1969) showed that, without correction, the probability of a type I error rate is seriously inflated when repeated significance tests are applied at a nominal level. Moreover,

the stopping rule in itself biases the final p-values. They are not uniformly distributed under the null hypothesis (Chang *et al.*, 1995). Several authors have proposed specific significance levels $\alpha' < \alpha$ for each interim analysis, (given that the total maximum number of analyses is known), such that the overall significance level α remains inferior to the chosen one, for example 5%. Levels of Peto *et al.* (1976) are very stringent at the first analysis in order to keep a significance level of 0.05 at the last analysis. There are very few chances to conclude the analysis prematurely. Later, Pocock (1977) and O'Brien and Fleming (1979) gave levels which are still widely used. Contrary to Pocock who gives constant levels, O'Brien and Fleming give slowly increased levels. Thus, it is less likely with the boundaries of O'Brien and Fleming to conclude differently from the conclusion which would have been hold in the case of a unique analysis. Pocock's approach allows for an earlier termination of the study than O'Brien and Fleming. But as pointed out by Sébille and Bellissant (2003), one drawback of Pocock's method is that the expected number of subjects required to conclude can be substantially increased when compared with the sample size required by the single-stage design of the same power.

These methods were generalized by Lan and DeMets (1983), who proposed a method where the number of analyses did not need to be fixed in advance. Their method, called α-spending, uses an increasing function $\alpha^*(t)$ with $\alpha^*(0)$, $\alpha^*(1) = \alpha$, where t is the fraction of the total information available at each analysis. Several functions can be found in the literature (Sébille and Bellissant, 2003): $\alpha^*(t) = \alpha ln(1 + (e-1)t)$ as an approximation of Pocock's significance levels, $\alpha^*(t) = 2 - 2\Phi(z_{\alpha/2}/\sqrt{t})$ as an approximation of O'Brien and Fleming's significance levels (where Φ is the standard normal distribution and $z_{\alpha/2}$ is the $\alpha/2$-fractile of $\mathcal{N}(0,1)$). Given the α-spending function, people define boundaries $\{b_1, \ldots, b_k\}$ such that the experiment is stopped at the kth interim analysis if the statistic $|S(k)|$ exceeds a chosen boundary value b_k. Under H_0,

$$P_0\{|S(1)| \leq b_1, \ldots, |S(k-1)| \leq b_{k-1}, |S(k)| > b_k\} = \alpha^*(t_k) - \alpha^*(t_{k-1}) = \pi_k \tag{3.1}$$

where π_k is the probability to stop the experiment at the kth analysis (Lee, 1994).

$$\pi_1 + \ldots + \pi_k = \alpha$$

Thus, $\alpha^*(t_k)$ is the probability to cross a boundary at or before the kth analysis (Spiessens *et al.*, 2000).

3.1.2 Calculation of b-values for Student distributions

Concerning the implementation of these methods, the α-spending methods are implemented in the function ldBands belonging to the R package Hmisc. It seems, however, that a normal distribution is always assumed to calculate b-values $b_k (1 \leq k \leq K)$ (with K the maximum number of analyses), which is reasonable in clinical trials but less meaningful for differential analysis in gene expression studies. As far as the microarrays are concerned, it is more usual to use the Student distribution because most of the time, very few samples are available. We looked in the literature and did not find any relevant paper suggesting a method to calculate b-values when the statistic follows a Student distribution. That is why we developed R-code ourselves to quickly calculate b-values for Student distributions by simulation.

Our approach is based on equation 3.1. To calculate the values of the boundaries $b_k (1 \leq k \leq K)$ for Student test statistics, we choose
$\Rightarrow$ K the number of analyses which will be performed
$\Rightarrow \pi_k (1 \leq k \leq K)$ with alpha spending methods (O'Brien and Fleming, 1979; Pocock, 1977)
$\Rightarrow n$ the number of observations by time and a large number of simulations

- For each simulation, we generate $2n$ $N(0,1)$
- We calculate $S(1)$ and keep it if $|S(1)| \leq b_1$
  with $b_1 = (1 - \pi_1/2)$- quantile of t_{2n-2}
- At each step, if $S(k-1)$ is kept, we generate $2n$ $N(0,1)$ and bind new data with precedent simulated data
- We calculate S(k) and keep it if $|S(k)| \leq b_k$
  with $b_k = (1 - \frac{\pi_k}{2(1-\sum_{j=1}^{(k-1)} \pi_j)})$ - empirical quantile

Once these b-values were calculated, we checked that with a large number of observations at each timepoint, they were close to b-values calculated initially by the function ldBands when normality was assumed.

For example, with $K = 3$ analyses, $n = 50$ observations by time, 100000 simulations, Pocock's method, b-values for Normal distributions are $b_1 = 2.279$, $b_2 = 2.295$, $b_3 = 2.296$ and b-values for Student distributions are $b_1 = 2.316$, $b_2 = 2.305$, $b_3 = 2.302$. In this case, we would suggest to use the original Pocock value of 2.3.

On the contrary, with a few number of observations by time, b-values are different, which shows the importance to calculate special b-values for

Student statistics. With $K = 4$ analyses, $n = 5$ observations by time, 100 000 simulations, O'Brien and Fleming method, b-values for Normal distributions are $b_1 = 4.333$, $b_2 = 2.963$, $b_3 = 2.359$, $b_4 = 2.014$ and b-values for Student distributions are $b_1 = 9.783$, $b_2 = 3.429$, $b_3 = 2.503$ and $b4 = 2.085$.

3.1.3 P-value combination

Another problem in sequential analysis is that there might be variance heterogeneity across the different stages of the analysis. This issue has led to meta-analysis methods being used in this context. Lehmacher and Wassmer (1999), instead of combining expression values up to the kth analysis and calculating boundaries for the test statistic, suggest to combine p-values from the different analyses. They use the test statistic that results from the inverse normal method of combining independent p-values (Hedges and Olkin, 1985)

$$\frac{1}{\sqrt{K}} \sum_{k=1}^{K} \Phi^{-1}(1 - p_k) \tag{3.2}$$

Lehmacher and Wassmer suggest then to take the classical group boundaries for the test statistic. The advantage of their method is that they do not have the problem of unknown variances and like in our previous method, they can have a Student distribution to calculate interim p-values. Their method is more general as they can use any distribution they want without calculating new b-values. They also allow for unequal samples sizes. In fact, combining p-values had already been proposed by Bauer and Kohne (1994), who focused on a two stage design and used Fisher's combination method:

$$S = -2 \sum_{k=1}^{K} ln(p_k) \sim \chi^2_{2K}$$

They did not use classical boundaries but calculated new boundaries for up to two analyses.

These methods that combine p-values and allow modification of the sample size of the second stage based on the predicted power of the trial at the end of the first stage are now called adaptive designs. An overview of such designs is given in Schäfer *et al.* (2006).

We did not investigate Bauer and Kohne (1994) method as Fisher's combination method requires to separate under and over-expressed genes.

The important question was then to know if there was a loss (and if yes, if it was big) of information when combining p-values by the inverse normal method rather than expression values. That is what we investigate in the next section.

3.2 Application to a simulated data set

We simulated one data set of 3000 genes with 100 over-expressed genes. The method and the parameters of simulation are given in Marot and Mayer (2009) (see section 3.3).

3.2.1 Genes studied separately

At first, we decided to make a gene-by-gene sequential analysis with a classical t-test, without any shrinkage or correction for multiple testing. Each gene was studied separately. We compared the two approaches described previously. The first method ('Student b-values') combines expression values up to the k^{th} analysis and calculates boundaries for the test statistic to stop the experiment before the end scheduled. The second one ('p-value combination') combines p-values from the different analyses. For this last method, as Lehmacher and Wassmer (1999) pointed out, it is necessary to use one-sided p values for both one-sided and two-sided cases to avoid directional conflicts, i.e. the case that opposite effects at previous stages may lead to the rejection of H_0 at some stage k of the procedure. For example, if a gene is a bit over-expressed at the first stage and a bit under-expressed at the second stage, then its test statistic is big at the first stage, small at the second one and there is no overall effect. If p-values are two sided, then p-values from both first and second stage are small and if they are combined they can lead to a rejection of the null hypothesis. If they are one-sided, one will be big whereas the other one will be small and this will avoid a spurious conclusion.

We compared in table 3.1 the 'Student b-values' and 'p-value combination' methods on a dataset where we added 5 replicates per time and performed 4 analyses. We counted the number of genes which would have been declared significant at each stage. In this analysis, as it was performed gene by gene, once the gene was declared significant, it was kept significant without checking if it was still significant in the following stage.

Both methods seem to be correct and almost all significant genes are found from the second stage. There is no loss of True Positives (TP) because of a combination of p-values rather than expression values. As we do not correct for multiple testing, we expect 5/100*3000=150 False Positives (FP) We see that both methods have a number of FP close to 150, a bit higher for the first one (Student b-values) and a bit smaller for the other one. Other simulations were performed and similar results were obtained.

As the second method performed very well, we decided to only use it in the following as it is more convenient for microarray data sets. Indeed, contrary to the first method, it does not raise difficult questions for normal-

Table 3.1: Results for 4 times with 5 observations per time - 3000 genes, 100 over-expressed, Pocock's boundaries

stage		1	2	3	4
Student b-values	TP	78	92	95	98
	FP	56	95	131	160
	FDR	0.42	0.51	0.58	0.62
p-value combination	TP	76	91	96	98
	FP	49	86	113	139
	FDR	0.39	0.49	0.54	0.59

ization. Combining expression values requires to normalize after each stage all data together. However, data obtained from different hybridizations are difficult to combine. Would we have to normalize each stage separately and then normalize them overall? The question would be how to normalize correctly without losing too much information. Another advantage of p-value combination is the possibility to use any desired modified t-test to test the differential expression.

As Victor and Hommel (2007) pointed out, controlling the FDR in adaptive designs has been considered in only few publications so far. They suggest to use the explorative Simes procedure (Simes, 1986). This requires an a priori on the number of differentially expressed genes N_0. Then, instead of using directly α to calculate the boundaries for sequential analyses, they calculate $\alpha_{Simes} = \frac{N_0}{N}\alpha$ with N the total number of tests. We thus calculated Pocock's boundaries with this new α_{Simes} and combined p-values by the inverse normal method. With an a priori of 100 differentially expressed genes, $\alpha = 0.05$, $\alpha_{Simes} = 100/3000 * 0.05 = 0.0016667$. We made an analysis with the same data as previously, keeping 4 analyses and 5 observations by time. Results are presented in table 3.2.

Table 3.2: Results for 4 times with 5 observations by time - 3000 genes, 100 over-expressed, FDR threshold 0.05, Pocock's boundaries

stage	1	2	3	4
TP	42	74	89	92
FP	1	1	2	4
FDR	0.02	0.01	0.02	0.04

We notice that the FDR is controlled and there is still a high number of true positives declared from the second stage. The major drawback of this method is that an a priori on the number of differentially expressed genes is needed.

To have a less conservative method than the Simes procedure, it would have been interesting to investigate the definition of the global p-value of Victor and Hommel (2007). They generalised the definition of Brannath *et al.* (2002) which was the probability that a p-value combination from the first and second stage would occur that was at least as extreme as the combination which was observed. Nevertheless, using their method required one decision by gene as all the analyses we did previously. At this stage, we were more interested in investigating further in one decision for the whole analysis rather than one decision gene by gene. Indeed, in microarray experiments, the decision about stopping or continuing the experiment has to be made for all genes simultaneously.

3.2.2 Multi-dimensional analysis

There are two directions to increase the dimensionality of an analysis. Either the number of variables or measurements for one individual are increased or the number of null hypotheses tested is higher than two. Clinical trials have been a wide field of research for both. We had only an interest in the number of hypotheses tested but more information about multivariate or longitudinal sequential analyses can be found in Wei *et al.* (1990), Spiessens *et al.* (2000), Todd (2007). In clinical trials, multiple null hypotheses simultaneously tested are often called 'multiple endpoints'. Tang *et al.* (1989) showed that a sample size based on multiple endpoints is smaller than the one based on any single endpoint when there are multiple endpoints. However, their situation is not directly applicable as they search to show that globally, treatment A is different from treatment B by comparing k multiple correlated endpoints. As far as we were concerned, we also wanted one decision at the end but we could not wait that our first condition would be different from the second as most of the genes are not differentially expressed. Another approach to multiple endpoints was given by Kieser *et al.* (1999). They considered the situation of a priori ordered hypotheses to end the experiment. As in the previous section, even if they considered multiple endpoints, they could stop for one endpoint separately from the other ones. Concerning the multiple testing when no ordering is possible, they controlled the Family Wise Error Rate (probability of making one or more false discoveries) which is more conservative than the False Discovery Rate. More recently, Xiong *et al.* (2005) proposed an intersection-union test which gives the minimum statistical power from the

individual tests. Their null hypothesis is the union of all endpoint-specific null hypotheses from the multiple endpoints. The corresponding alternative hypothesis is the intersection of all alternative hypotheses. Once more, this method was not applicable in our situation as we did not know a priori which genes we wanted differentially expressed. We could neither have only null hypotheses with non differentially expressed genes. The stopping criterion we were looking for was a summary of significance of all individual tests but without knowing a priori which ones were significant.

On the same simulated data set as previously, we studied how continuing the experiments for all genes influenced the outcome. The bottom of table 3.3 gives the same results as table 3.1 in order to compare a sequential analysis where differentially expressed genes were kept from one stage to the following one with a sequential analysis where they are tested at each stage.

Table 3.3: Results for 4 times with 5 observations by time - 3000 genes, 100 over-expressed, Pocock's boundaries

stage		1	2	3	4
All genes tested at each stage	TP	76	91	96	98
	FP	49	52	51	62
Differentially expressed genes kept at each stage	TP	76	91	96	98
	FP	49	86	113	139

In the case where all genes were tested at each stage, we calculated the number of differentially expressed genes which were not declared at the previous stage or at the following stage. Thus, 34 (resp. 31, 25) genes which were declared at the first (resp. second, third) stage were not anymore at the second stage (resp. third, fourth) but 52 (resp. 35, 38) other genes were found differentially expressed. We notice that on this example, only false positives disappeared from a stage to the following one. The consequence was that when they were kept, there were many more false positives. The number of false positives at the final stage was, however, far smaller than the expected number of 145. As the stopping rule was made from the summary of 3000 genes without stopping for each individual gene, one could ask if it was not too much to correct both for multiple testing and bias implied by the stopping rule.

It is reasonable to think that as the stopping rule is based on the distribution of all p-values or test-statistics, the more genes there are, the less individual value is biased by this rule. The following paper presents histograms

illustrating this idea. Assuming that the stopping decision and the final p-value are nearly independent, the problem of sequential analysis is simplified and classical methods to calculate boundaries are not needed. This leads to a new definition of sequential analysis for microarray data. That is what we developed in the following paper (Marot and Mayer, 2009).

3.3 Sequential analysis for microarray data based on sensitivity and meta-analyses

Statistical Applications in Genetics and Molecular Biology

Volume 8, *Issue* 1 2009 *Article* 3

Sequential Analysis for Microarray Data Based on Sensitivity and Meta-Analysis

Guillemette Marot* Claus-Dieter Mayer†

Sequential Analysis for Microarray Data Based on Sensitivity and Meta-Analysis*

Guillemette Marot and Claus-Dieter Mayer

Abstract

Motivation: Transcriptomic studies using microarray technology have become a standard tool in life sciences in the last decade. Nevertheless the cost of these experiments remains high and forces scientists to work with small sample sizes at the expense of statistical power. In many cases, little or no prior knowledge on the underlying variability is available, which would allow an accurate estimation of the number of samples (microarrays) required to answer a particular biological question of interest. We investigate sequential methods, also called group sequential or adaptive designs in the context of clinical trials, for microarray analysis. Through interim analyses at different stages of the experiment and application of a stopping rule a decision can be made as to whether more samples should be studied or whether the experiment has yielded enough information already.

Results: The high dimensionality of microarray data facilitates the sequential approach. Since thousands of genes simultaneously contribute to the stopping decision, the marginal distribution of any single gene is nearly independent of the global stopping rule. For this reason, the interim analysis does not seriously bias the final p-values. We propose a meta-analysis approach to combining the results of the interim analyses at different stages. We consider stopping rules that are either based on the estimated number of true positives or on a sensitivity estimate and particularly discuss the difficulty of estimating the latter. We study this sequential method in an extensive simulation study and also apply it to several real data sets. The results show that applying sequential methods can reduce the number of microarrays without substantial loss of power. An R-package SequentialMA implementing the approach is available from the authors.

KEYWORDS: microarrays, sequential analysis, meta-analysis

---

*Guillemette Marot's PhD is supported by INRA, Département de Génétique Animale and INRA, Département de Physiologie Animale et Systèmes d'Elevage. Claus Mayer's work is funded by the Scottish Government Rural and Environment Research and Analysis Directorate (RERAD).

1 Introduction

Sequential methods and adaptive designs have a long standing tradition in clinical trials and overviews may be found in Lee (1994), Spiessens *et al.* (2000), Sébille and Bellissant (2003), Schäfer *et al.* (2006), Todd (2007). These methods are characterised by interim analyses at pre-defined stages and a stopping rule which determines at each stage whether to stop or continue sampling. Naturally, this stopping rule is based on data collected up to the current stage only. Sequential methods thus have the potential to reduce the required sample size. This is important in clinical trials or animal experiments where it might be unethical to collect more samples than necessary, but it is also an interesting feature in situations where either collecting or analysing a sample is very expensive or time-consuming.

Microarray experiments are a typical example of the latter situation. Although prices have come down, commercial microarrays remain costly, whereas home-spotted two-colour arrays might be cheaper but typically involve a very time-consuming scanning process. Microarray experiments also often have technical limits with respect to the number of samples that can be analysed simultaneously. For example a hybridisation chamber will only be able to accommodate a limited number of arrays. This means that not all samples might be available for data analysis at the same time but that they arrive in a staggered fashion. Rather than waiting until all data are available it seems natural to analyse the data present at each stage and potentially stop the process once the results fulfil certain criteria, which we will discuss in detail later. Hence microarray analysis lends itself naturally to the application of sequential methods.

Even though keeping costs small is desirable, the main objective of any biological experiment is to detect effects, which in a statistical testing framework corresponds to the notion of statistical power. Since sample size reduction automatically decreases the power of an experiment it is clearly important to find a balance between these two contrasting aims.

The classical way of finding this balance is to perform a sample size/power calculation prior to the experiment. However, this calculation requires prior knowledge of effect sizes and variability that is often lacking in real life situations. The sequential approaches that we discuss can be interpreted as designing an experiment that contains its own pilot study (or studies) and then utilises the information obtained from these pilot studies to update the power calculation.

Classical sequential methods deal with the analysis of one variable only. A review on generalisations to the multivariate case can be found in Lee (1994).

That paper, however, focuses on a repeated measurement situation where the same variable is measured at different timepoints. Recently, Victor and Hommel (2007) studied the use of interim analyses in high dimensional cases. A fundamental difference between their experimental structure and a microarray experiment is that they consider a case where each variable has an individual stopping rule, whereas in a microarray experiment the sampling process has to be stopped or continued for all genes simultaneously. Due to this difference, the problems in developing a sequential method turn out to be very different from the ones studied in Victor and Hommel (2007). In fact the sequential strategies for microarray experiments are relatively simple. This is due to the fact that stopping rules in this context do not depend on specific single genes but on the distribution of statistics or p-values across all genes. As a result, the type 1 error/false disccovery rate is not inflated, one of the main problems in the classical situation.

We mainly discuss stopping rules based on sensitivity, i.e. at each stage we estimate the percentage of truly differentially expressed genes among those declared significant. We propose to stop the experiment if this estimated sensitivity exceeds a pre-defined threshold. This seems a sensible approach in situations where we expect only a small to moderate number of genes to show changes. In cases with many differentially expressed genes, sensitivity remains low unless the sample size is large and in these cases we propose to use the estimated number of true positives that have been detected.

Because sensitivity, sometimes also called the expected discovery rate (EDR, cf. Gadbury *et al.* (2004)), plays such a central role in the proposed sequential design, a considerable part of our paper discusses a number of different sensitivity estimators.

In addition to finding a stopping rule, the second major issue in our approach is how to combine the data from the different stages of the experiment. One possibility would be to simply re-analyse the complete data set after each stage. Since some microarray normalisation tools like for example *GCRMA* (Wu *et al.*, 2004) use information across all arrays for normalisation, this would typically mean re-normalising the arrays at each stage, which would lead to inconsistency in the data. Another problem with this approach is that there might be stage effects (for example arrays hybridised simultaneously tend to show higher correlations) which would both affect normalisation and subsequent analysis. For these reasons we propose a meta-analysis type approach, in which data from each stage are analysed separately and only the p-values from each stage are combined in the end.

In the next section we will give a more detailed description of the sequential analysis we propose, chronologically following the different steps from planning

of the experiment to sensitivity estimation based on combined p-values. Section 3 presents simulations of certain aspects of our methods as well as the overall effects of the sequential approach and also an application to two real data sets. We summarise our findings and give an outlook on further research topics in section 4.

2 Methods

We now describe our sequential approach in more detail. In this paper we only consider the standard two-sample problem, i.e. that we have samples being taken under two different experimental conditions or from two groups of phenotypes. In principle, other more complex designs like the k-sample problem, more-factorial situations etc. can also be treated in a sequential fashion and we will touch upon this briefly in our discussion. We assume these samples to be independent biological replicates. Further, we assume that the microarray technology used will give us a single expression value per gene and sample, which could either be an absolute measurement of gene expression obtained from a single channel microarray (for example an Affymetrix chip) or a relative measurement of gene expression with respect to a common reference on a two-colour array. Let G denote the number of genes (probes) spotted on the array, which are indexed by $g = 1, \ldots, G$. We want to allow for a maximum of K interim analyses (indexed by k) with sample sizes $n_1(k), n_2(k)$ in groups 1 and 2 at stage k and with $n(k) = n_1(k) + n_2(k)$ denoting the total number of samples for stage k. The maximum sample size will thus be $N = \sum_{k=1}^{K} n(k)$.

A sequential analysis plan will then consist of the following steps

1. Designing the experiment: How many arrays N will be maximally used? What is the choice of K and $n_1(k), n_2(k)$?

2. Analyses of the current stage: This will produce p-values $\tilde{p}_g(k)$ for gene g and stage k, which corresponds only to the $n(k)$ samples analysed at this stage.

3. Analyses up to stage k: This will produce p-values $p_g(k)$ that are based on all $\sum_{j=1}^{k} n(j)$ samples that are available so far.

4. Stopping decision: Based on the current p-value vector $(p_1(k), \ldots, p_G(k))$ we decide to proceed with stage $k+1$ or to stop the experiment. If $k = K$ the experiment is stopped anyway.

5. Final analyses: Once the experiment has been stopped at a stage $k \leq K$, we have to make a final decision for each gene whether we declare it to be differentially expressed or not. This decision should take the multiple testing problem into account, which arises from the fact that thousands of decisions have to be made simultaneously.

Below, we will discuss different options for each of the 5 steps. These options are mainly already established methods within microarray statistics. Although we have preferences for certain methods ourselves, they may be chosen flexibly. The following list gives an overview for the methods we discuss for the five steps explained as well as some of our personal recommendations.

1. Power calculation tools for microarrays can be used to determine a maximal number of samples/arrays. Examples are *PowerAtlas* (Page *et al.*, 2006) or *sizepower* of Lee and Whitmore (2002), see also section 2.1. We recommend at least 4 arrays per group for the first stage (note that we consider smaller sample sizes for illustration purposes in our examples though).

2. We suggest to use the Bioconductor *limma* package Smyth (2004) to analyse the arrays within a stage, cf. section 2.2.

3. We use the inverse normal method to combine p-values from different stages, see section 2.3.

4. We propose a stopping criterion based on either an estimator of the sensitivity (defined in Formula 2.4.1) or on the estimated number of true positives. The estimation methods are listed in Table 1 and discussed in detail in section 2.4. Based on our observations we cannot give a clear recommendation on which method to use. If sensitivity is used for the stopping criterion the choice of threshold is up to the user and the aim of the experiment but we think that values in the range of 60% to 90% will be typically used. In cases with many differentially expressed genes, sensitivity will only increase very slowly and in such a situation a criterion based on the estimated number of true positives seems more appropriate.

5. We suggest to use the Benjamini-Hochberg method to control for false discovery rate, when selecting significant genes.

2.1 Experimental Design

The issue of power and sample size calculations for microarray data has already been studied by several authors, cf. Lee and Whitmore (2002), Page *et al.* (2006), Liu and Gene Hwang (2007). Some tools are available in packages like the R *sizepower* library of Lee and Whitmore (2002) or as web-based software like *PowerAtlas* (Page *et al.*, 2006). However, typically these calculations require prior knowledge of effect sizes and variability, which may be obtained from pilot studies. *PowerAtlas* (Page *et al.*, 2006) is interesting in this respect since it allows the user to find a microarray data set from the GEO-database (Edgar *et al.*, 2002) that is similar to the planned experiment and bases the power calculation on this experiment.

These tools could be used to estimate a maximal total sample size initially and *PowerAtlas* also allows using the data collected up to a given stage to predict power for later stages and thus could be used to choose the sample sizes for these latter stages adaptively. In practice, however, this maximal number is often determined by external considerations, like the budget for the study or the number of available samples. Similarly the number of arrays that will be analysed in each interim analysis is often not chosen by the scientist but depends on how many hybridisations are possible simultaneously. In general we suggest using a balanced design, i.e. to choose equal sample sizes $n_1(k) = n_2(k) = n(k)/2$ at each stage, since this tends to maximise power. Note that for a statistical analysis, at least 2 replicates per group are needed at any stage. We would recommend larger sample-sizes per stage though as the methods discussed in the following assume that we obtain valid p-values from the data collected at each stage and it is known that p-values tend to be not exact for very small sample-sizes. We recommend a sample-size of at least 4 per group in particular for the first stage.

2.2 Analysis of the current stage

Once the data from the hybridisations at stage k are available for analysis, we suggest normalising and analysing them with standard methods developed for microarrays. If the normalisation uses information across arrays (e.g. GCRMA Wu *et al.* (2004)), only the $n(k)$ hybridisations from the current stage will be normalised simultaneously. This avoids re-normalising arrays from previous stages, which might lead to conflicting results. It also takes into account that arrays hybridised at different stages might have different properties, which makes a joint normalisation questionable anyway.

When testing for differences between the two groups we suggest using the

moderated t-test offered in the Bioconductor package *limma*, Smyth (2004). Alternative methods are possible too as long as they give a uniform p-value distribution for non-differentially expressed genes. Note that the moderated t-test in limma has a larger number of degrees of freedom than the classical t-test and thus is more stable than a classical t-test. It assumes normally distributed data in the calculation of p-values, an assumption that we have found to hold reasonably well in many cases, see also paragraph 6.2.2 in Wit and McClure (2004), who state "..., it is our experience that the misspecification made by using a normal approximation is typically negligible". The limma t-test is widely used for microarray data with small sample-sizes and we have found it to produce sufficiently valid p-values in most cases. Still it is recommendable to study p-value histograms for diagnostic purposes for each stage. Examples of "correct" and "uncorrect" histograms of p-values are given in Page *et al.* (2006). We find that "strange" p-value distributions are usually not caused by small sample sizes or lack of normality but by sources of variation that are either unknown or have not been taken into account by the statistical model.

2.3 Combining different experiments

2.3.1 Controlling the false discovery rate

One fundamental problem in sequential analyses is that the results obtained once stage k is complete have to be analysed conditionally given that the experiment was not stopped at stages $1, \ldots, k-1$. As a consequence of this, the p-value distribution is no longer uniform under the null hypothesis in latter stages, c.f. Chang *et al.* (1995). To illustrate this, let us consider a univariate sequential study with maximally two stages, that yields p-values p_1 from stage 1 and p_2 from the cumulative analysis of stages 1 and 2. Assume that the study is stopped if the first p-value is below a significance level α and continues otherwise. If p denotes the final p-value of this procedure, we can write its distribution as

$$P(p \leq t) = P(p_1 \leq t | p_1 \leq \alpha)\pi + P(p_2 \leq t | p_1 > \alpha)(1-\pi) \qquad (1)$$

with $\pi = P(p_1 \leq \alpha)$. We see that this is a mixture of two distributions, none of which will be uniform even if p_1 and p_2 are uniformly distributed. Note that the two p-values are not independent since the data from stage 1 are present in both.

In a microarray situation we have p-values $p_g(k)$, for each gene g and stage k. The stopping rules we consider always depend on the empirical distribution of p-values only and do not utilise the knowledge of which p-value corresponds

to which gene. As the number of genes increases, the empirical distribution of all p-values contains less and less information about each individual one and thus the dependence between the p-values and the stopping decision vanishes asymptotically, e.g. the correlation of a single p-value with the empirical distribution function converges to zero at a rate of $1/\sqrt{G}$ (see Appendix) .

To see how this asymptotic result holds for different values of G, we simulated p-values from an analysis of normally distributed data with known variances, i.e. we chose

$$p_{1g} = 1 - \Phi(S_{1g}), \quad p_{2g} = 1 - \Phi((S_{1g} + S_{2g})/\sqrt{2}), \tag{2}$$

where the S_{ig} were independent standard normally distributed under the null hypothesis and had mean 1 under the alternative. As a stopping rule we chose to end the experiment if there was at least one significant result at the first stage after adjusting p-values by the Benjamini-Hochberg (BH) method. Figures 1-3 show histograms of the final p-value for one particular gene in the situation where there are 1, 5 or 1000 genes on the array. This particular gene as well as all other genes were assumed to follow the null hypothesis, i.e. being non-differentially expressed.

Figure 1: Distribution of the final p-value after a two-stage sequential analysis with $G = 1$ gene on an array (Distribution estimated from 1000 simulations).

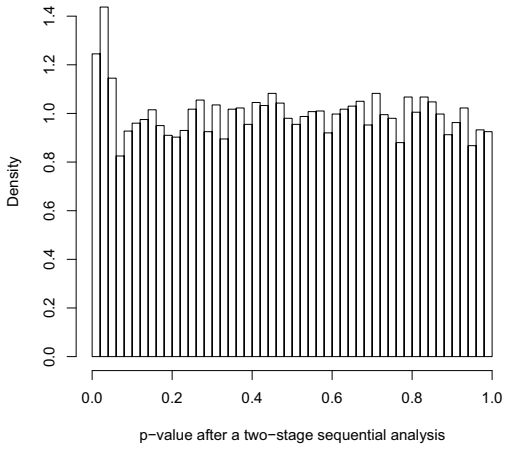


Figure 2: Distribution of the final p-value after a two-stage sequential analysis with $G = 5$ genes on an array (Distribution estimated from 1000 simulations).

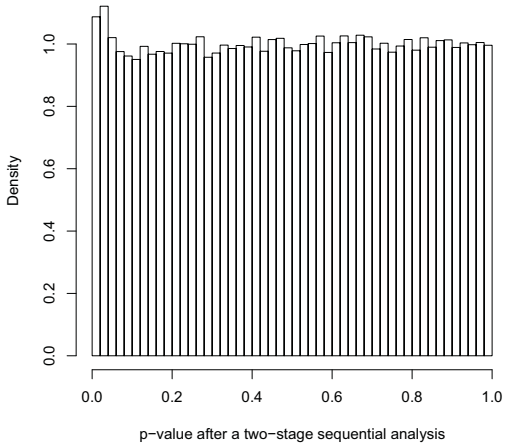
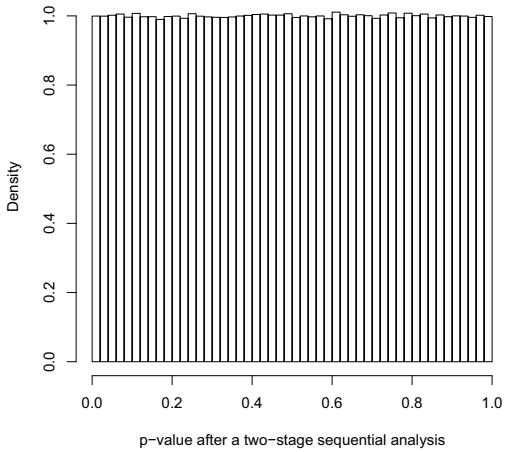


Figure 3: Distribution of the final p-value after a two-stage sequential analysis with $G = 1000$ genes on an array (Distribution estimated from 1000 simulations).



We see that for $G = 1$ and $G = 5$ genes there is a clear bias towards small p-values, whereas for $G = 1000$ (which is a small number in a microarray context), this bias is not detectable any more. The simulated situation is a worst-case scenario in the sense that the presence of differentially expressed genes will reduce the bias of the p-value distribution of non differentially expressed genes. Also, a stricter stopping rule, e.g. demanding more than one significant gene, will result in fewer experiments being stopped at the first stage and thus reduce bias. Only increasing the number of interim analyses would increase the bias but other simulations (not presented here) showed that in realistic situations with 3-5 stages and more than 1000 genes, the effect remains negligible.

The fact that the stopping decision and the final p-value are nearly independent simplifies the sequential design a lot. In the classical univariate case it is difficult to control the type-I-error in a sequential test. One approach to deal with this issue is the "α-spending methods" discussed by Lan and DeMets (1983). Victor and Hommel (2007) developped a method to control the false discovery rate in an interim analysis approach, where the decision to stop the experiment is made individually for each variable. As their paper shows, this situation is considerably more complex than the univariate situation. In microarray analysis, however, the experiment is stopped or continued for all genes simultaneously and thus controlling the false discovery rate is straightforward. We can simply apply the Benjamini-Hochberg method to the final p-value vector.

2.3.2 Combining p-values from different stages

Microarray data are influenced by many unwanted sources of variation and ideally all other parameters apart from the experimental conditions of interest should be kept as constant and homogeneous as possible. In a sequential study, data obtained at different stages will inevitably have slightly different characteristics. If these effects are mild they possibly can be eliminated by joint normalisation of all data or be taken into account by including a stage effect in the statistical analysis. In more severe cases though, where data from different stages differ in more complex ways than a simple shift or change of scale of the data distribution, a joint analysis of all data can be challenging.

This is related to the statistical field of meta-analysis, which investigates approaches to combine data from different studies or centres and to model the inter-study variation. Choi *et al.* (2003) were among the first authors to address this problem in a microarray context and some of their methods are implemented in the Bioconductor library *GeneMeta* described by Gentleman

et al. (2006). Following their work, Conlon *et al.* (2007) compared Bayesian meta-analysis models, whereas Stevens and Doerge (2005) studied the special case of combining Affymetrix data from different studies. In all these approaches, the expression values themselves are combined.

An alternative approach to meta-analysis is to combine the p-values obtained from different studies instead of combining the expression values themselves. Unlike the full data, p-values are always comparable across different studies and they implicitly contain information about within-study variability. The disadvantage is that using p-values only means a loss of information and thus might cause a reduction in power. We regard a p-value combination approach as a conservative method, that will be less affected by between study differences. We consider two different methods. The first is Fisher's method, which combines the p-values for gene g up to stage k in this way:

$$S_g = -2 \sum_{j=1}^{k} \ln(\tilde{p}_g(j)). \tag{3}$$

Under the null hypothesis S_g follows a chi-square distribution with $2k$ degrees of freedom. This combination method has been used before for microarray statistics by Hess and Iyer (2007) but in the completely different context of combining p-values obtained for individual probes within a probeset for Affymetrix data.

The second method considered here is the inverse normal method, cf. Hedges and Olkin (1985):

$$S_g(k) = \frac{1}{\sqrt{k}} \sum_{j=1}^{k} \Phi^{-1}(1 - \tilde{p}_g(j)) \tag{4}$$

Under the null hypothesis $S_g(k)$ follows a standard normal distribution. This combination of p-values was suggested for group sequential trials by Lehmacher and Wassmer (1999) and yields a statistic that follows a standard normal distribution under the null hypothesis. As Lehmacher and Wassmer (1999) point out, it is necessary to use one-sided p-values $\tilde{p_g}$ to avoid contradictory results from different stages leading to significant findings. For the inverse normal method, consistently low (high) p-values across all stages will lead to a high (low) value of the combination statistic. An overall two-sided p-value can then be obtained by

$$p_g(k) = 2(1 - \Phi(|S_g(k)|)), \tag{5}$$

whereas Fisher's combination method demands a separate combination of one-sided p-values for up or down-regulation. Another advantage of the inverse

normal method is that the combination of the p-values is identical to summing up the original test-statistics for the case of normally-distributed statistics. When testing for differences between normal means with known common variance, this method thus yields the optimal test statistic, if the sample sizes are the same at each stage. A weighted variant of the inverse normal method using

$$S_g(k) = \sum_{j=1}^{k} w_j \Phi^{-1}(1 - \tilde{p}_g(j)) \tag{6}$$

with

$$w_j = \sqrt{\frac{n(j)}{\sum_{i=1}^{k} n(i)}}$$

maintains this property also for unequal sample sizes at different stages. Under the null hypothesis $S_g(k)$ follows a standard normal distribution. In contrast to the unweighted case and Fisher's method this combination method gives more weight to stages with more samples, which seems to be sensible, as long as we assume that there is no change in variance between different stages. Hedges and Olkin (1985) discuss such weighted inverse normal approaches in more detail. Loughin (2004) compares different p-value combination methods in a simulation study and states that the (unweighted) inverse normal method works well in cases where the evidence against the null-hypothesis is spread equally across the different studies. This is a reasonable assumption in our application where instead of different studies we are dealing with different stages of the same experiment.

To decide which genes are significant after stage k, we propose to apply the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) to adjust the p-values $p_g(k)$ for multiple testing. For the remainder of the paper a gene will be regarded *as differentially expressed* or *significant*, if its adjusted p-value is below a threshold α, for example $\alpha = 5\%$.

2.4 Stopping Criteria

The decision at which stage to stop the experiment is the most crucial component of our sequential approach. Whether one considers the information collected up to a given stage as sufficient depends on the context and the aim of the study. We differentiate between two principal situations. In the first scenario the microarray experiment serves as an initial screening step to filter out the most differentially expressed genes, which will then be analysed in more detail in a follow-up experiment involving other technologies (e.g. quantitative

RT-PCR analysis). Often only a small number of genes can be followed up in this way, so that it seems natural to stop the experiment as soon as this number of differentially expressed genes has been found with high confidence.

In the second and probably more common situation, the experiment is conducted to obtain a global picture on which genes and pathways are changing between different experimental conditions. In this case it is not a fixed number of significant genes we are interested in, but the aim is to find a large proportion of all genes that are differentially expressed genes, i.e. to ensure a good sensitivity of our method. We thus suggest estimating the sensitivity based on the current set of p-values after each stage and stop the experiment if this estimate exceeds a user-defined threshold. The remainder of this section discusses different ways of estimating sensitivity, most of which are based on fitting mixture distributions to the histogram of observed p-values.

2.4.1 Estimation of sensitivity or EDR

The Expected Discovery Rate (EDR) was introduced to microarray analysis by Gadbury *et al.* (2004), but we will mainly follow the nomenclature of Pawitan *et al.* (2005) and refer to the EDR as *sensitivity*. For a proper definition of the EDR/sensitivity, we consider a collection of genes/p-values, which can be grouped in two ways: they can either correspond to the null hypothesis, i.e. be negatives or the alternative (positives); secondly they can be declared significant or not. As a result, we have four groups of genes, i.e. any gene will either be a true negative (TN), a false negative (FN), a false positive (FP) or a true positive (TP) and we will use the same abbreviations for the number of genes falling into these categories.

	Not a real effect	Real effect
Not declared significant	TN	FN
Declared significant	FP	TP

Note that here TN, TP, FP and FN are counts and not proportions. Also note that these 4 numbers are random variables, since the decision of declaring them significant depends on the data. The sum of these four figures is the total number of genes in the experiment, and we know the number of significant genes $TP + FP$ and the number of non-significant genes $TN + FN$, but in real life situations we do not know how many true positives and true negatives there are, so we will have to estimate these quantities from the data. Using the convention $0/0 := 0$ we define the false discovery rate (FDR) and expected discovery rate (EDR) as

$$FDR = E(\frac{FP}{TP + FP}) \tag{7}$$

and

$$sensitivity = EDR = E(\frac{TP}{FN + TP}). \tag{8}$$

Note that the word sensitivity and power are often used synonymously and will be used by us that way too at times to make the text more readable. In a strict sense this is a misuse of terminology since power refers to a single test for one gene, whereas sensitivity characterises the overall testing procedure for *all* genes. The estimation of this EDR or sensitivity requires the estimation of TP and FN. We consider different estimation approaches that have been proposed before.

Allison *et al.* (2002) modelled the p-value distribution as a mixture of $\nu + 1$ beta distributions on the interval $[0, 1]$. Gadbury *et al.* (2004) used this mixture to estimate the number of TP, TN and EDR. For a single set of p-values these estimates can be obtained by using the web-based poweratlas software (www.poweratlas.org). Since we used the programming language R for our simulations, we were not able to include this method in our simulation study.

A similar and simpler approach is the Beta Uniform Model (BUM) developed by Pounds and Morris (2003), which is based on the mixture of only two beta distributions and is available in a collection of R libraries called *OOMPA* (http://bioinformatics.mdanderson.org/Software/OOMPA/). The probability distribution function fitted by BUM is

$$f(p|a, \lambda) = \lambda + (1 - \lambda)ap^{a-1} \tag{9}$$

for $0 < p \leq 1$, $0 < \lambda < 1$, and $0 < a < 1$. Since the non-uniform part of this mixture is strictly positive even at $p = 1$ Pounds and Morris (2003) did not use the fitted mixture parameter $\widehat{\lambda}$ as an estimate for the proportion of non differentially expressed but the upper bound given by

$$\widehat{\pi}_{ub} = \widehat{\lambda} + (1 - \widehat{\lambda})\widehat{a}.$$

If h is a given p-value threshold, TP, TN, FP and FN can be estimated as

$$\widehat{TP} = G(\widehat{F}(h) - \widehat{\pi}_{ub}h) \tag{10}$$

$$\widehat{FN} = G(1 - \widehat{F}(h) - (1-h)\widehat{\pi}_{ub})$$
$$\widehat{FP} = G\widehat{\pi}_{ub}h$$
$$\widehat{TN} = G(1-h)\widehat{\pi}_{ub},$$

where G is the total number of genes and $\widehat{F}(h) = \widehat{\lambda}h + (1-\widehat{\lambda})h^{\widehat{a}}$ is the distribution function corresponding to the fitted density. Our sensitivity estimate then is

$$\widehat{EDR} = \frac{\widehat{TP}}{\widehat{TP} + \widehat{FN}}. \tag{11}$$

The p-value threshold h_0 we suggest to use is the one that controls the FDR according to the Benjamini-Hochberg rule. Note that the sum of estimated true positives and false positives does not equal the number of significant genes since the fitted distribution function is not identical to the empirical distribution.

As an alternative, we thus considered an approach that does not use a fitted mixture but only the empirical distribution and an estimate of the proportion π_0 of non differentially expressed genes. To estimate this proportion, we used the method of Langaas *et al.* (2005), which is implemented in the function convest in the Bioconductor package *limma* (Smyth, 2004). This estimator $\widehat{\pi_0}$ is based on a non parametric maximum likelihood estimation of the p-value density. We then used the previous equations replacing $\widehat{F}$ with the empirical cumulative distribution function and $\widehat{\pi}_{ub}$ by $\widehat{\pi_0}$. We will refer to this proportion based empirical approach as the "PE method".

Other mixture models were used by Efron (2004) and McLachlan *et al.* (2006). We would like to stress that most of these models were originally introduced to estimate the (local) FDR in a multiple testing situation, but will be used by us to estimate the EDR. Both do not fit a distribution to the p-values themselves but fit normal mixtures to transformed p-values. One of the advantages of such a transformation is that it allows to use a wide range of mixture fitting tools which have been specifically developed for normal distributions.

McLachlan *et al.* (2006) converted the two sided p-values to z-scores via $z = \Phi^{-1}(1 - p_{2sided})$ where Φ is the $N(0,1)$ distribution function. High z-scores correspond to small p-values and thus differentially expressed genes in this approach.

Efron (2004) used a similar transformation based on one sided p-values by defining $z = \Phi^{-1}(p_{1sided})$ (up to a sign this is identical to the p-value transformation used in our p-value combination approach in (4) and (6)). Here very high or very low z-scores both indicate differential expression.

In both methods the distribution of z-scores is modelled using a mixture

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z), \tag{12}$$

where π_0 denotes the proportion of non differentially expressed genes, f_0 is the density of transformed p-values under the null hypothesis and f_1 models the p-value distribution of differentially expressed genes.

Let ϕ denote the density of a normal distribution. McLachlan assumes that

$$f(z) = \pi_0 \phi(z; \mu_0, \sigma_0^2) + (1 - \pi_0)\phi(z; \mu_1, \sigma_1^2) \tag{13}$$

He suggests two different procedures: the "theoretical null procedure", which assumes $\mu_0 = 0$, $\sigma_0^2 = 1$, and the "empirical null procedure", which estimates μ_0 and σ_0^2 from the data. For estimation we used the EM algorithm as described in McLachlan *et al.* (1999). As an initial value $\pi_0^{(0)}$ for π_0, we chose $\pi_0^{(0)} = \widehat{\pi_0}$ estimated by the method of Langaas *et al.* (2005). The other initial parameters were derived based on $\pi_0^{(0)}$ using the equations given in McLachlan *et al.* (2006). We will refer to the two approaches as "NT" and "NE".

Efron (2004) earlier had proposed a more general method for his z-transformed one-sided p-values, which has been implemented in the *locfdr* package in Bioconductor. It offers the same two options for the null hypothesis distribution, i.e. f_0 can be modelled either as a standard normal distribution or as a general normal, where the parameters are estimated from the data (as above we will refer to these two options as the "theoretical" and "empirical" method), but in contrast to McLachlan's method though the second component f_1 is not specified but estimated non-parametrically. This obviously makes the estimation problem more complex and we refer to Efron (2004) for details. The two methods will be abbreviated as "LT" and "LE" below.

Note that in both methods the "empirical" option corresponds to a non-uniform distribution of p-values under the null-hypothesis. Efron (2004) discusses in more detail in what situations this might be advantageous. In this article we assume that the p-values obtained are exact or at least approximately exact, so that we would expect the p-value distribution to show only minor deviations from uniformity under the null hypothesis. As we show below all methods discussed will not only give us an estimate of sensitivity but also of the FDR. Under our assumption, a large difference between an FDR estimate and the nominal level controlled by the BH method indicates a problem with the FDR estimation method and thus the comparison between nominal and estimated FDR can be used to decide which estimation method should be used. We will come back to this point, when applying our method to real data sets in section 3.2.

Once the parameters and functions of these models have been estimated, let $\widehat{f} = \widehat{\pi_0}\widehat{f_0} + (1 - \widehat{\pi_0})\widehat{f_1}$ denote the estimated density. We can then estimate the posterior probability τ_0 of a gene with a transformed p-value z being non-differentially expressed as

$$\widehat{\tau_0}(z) = (\widehat{\pi_0}\widehat{f_0})/\widehat{f}(z). \tag{14}$$

Note that τ_0 is sometimes called the *local false discovery rate* ("*locfdr*"), although Efron refers to the upper bound f_0/f as *locfdr*. Based on this estimate we follow McLachlan *et al.* (2006) to obtain the following estimates:

$$\begin{aligned}
\widehat{TN} &= \sum_{g=1}^{G} \widehat{\tau_0}(z_g) I_{]c_0,\infty[}(\widehat{\tau_0}(z_g)) \\
\widehat{FN} &= \sum_{g=1}^{G} (1 - \widehat{\tau_0}(z_g)) I_{]c_0,\infty[}(\widehat{\tau_0}(z_g)) \\
\widehat{FP} &= \sum_{g=1}^{G} \widehat{\tau_0}(z_g) I_{[0,c_0]}(\widehat{\tau_0}(z_g)) \\
\widehat{TP} &= \sum_{g=1}^{G} (1 - \widehat{\tau_0}(z_g)) I_{[0,c_0]}(\widehat{\tau_0}(z_g))
\end{aligned} \tag{15}$$

The threshold c_0 here is chosen in such a way that the number of significant genes $FP + TP$ equals the number of genes being declared significant by the Benjamini-Hochberg rule. Note that the genes with $z_g > \widehat{\tau_0}(z_g)$ will not necessarily be identical to the ones declared significant by the BH method, if $\widehat{\tau_0}$ is not a monotonically decreasing function of the original two-sided p-value. In McLachlan's method this monotonicity can be affected if the variances of f_0 and f_1 are very different, whereas Efron's method will also be affected by a lack of symmetry of up and down-regulated genes. For all these methods the sensitivity (EDR) is estimated according to equation (11).

We remark that the list of estimation methods discussed here is not exhaustive since we primarily focussed on approaches that were available within R at the time of writing. An interesting alternative might be the approach of Robin *et al.* (2007), which, similarly to Efron's method, allows to fit a mixture, where one of the mixture components is known and the other one is not. Another possibility is to use not only the proportion estimate from Langaas *et al.* (2005), but make use of the non-parametric density estimation it is based on.

3 Results

Having described the sequential analysis approach and the options within it, we will now study our method applied to data. We start with several simulations to compare the different options within specific components of our approach. In a second step we demonstrate how the overall method behaves when applied to real microarray data. All simulations and calculations were performed using R 2.4.0.

3.1 Simulation study

We conducted two sets of simulations. Our first objective was to compare the different options for the stopping criterion, i.e. we compared the different sensitivity estimation methods described above. Secondly we studied the complete sequential approach, where we first compared the different meta-analysis approaches to combine p-values from different stages and then simulated a whole sequential study using one specific stopping criterion. This simulation allowed us to study a) whether there is a substantial loss in power by using a p-value combination approach instead of combining the original data, and b), what reduction in sample size a sequential approach might achieve.

3.1.1 Details of simulation strategy

We used the simulation strategy described in Delmar *et al.* (2005). In all simulations the number of genes was $G = 3000$. We simulated normally distributed gene expression values, where the parameters were calculated from the spleen data given in the R *Varmixt* library (Delmar *et al.*, 2005). Let $\mu_{1g}, \mu_{2g}, \sigma^2_{1g}, \sigma^2_{2g}$ be the empirical means and variances for gene g in the two different conditions of this real data set and let $\mu_g = (\mu_{1g} + \mu_{2g})/2$ denote the average of the means.

Differentially expressed genes were generated with $N(\mu_g, \sigma^2_{1g})$ for the 1st condition and $N(\mu_g + \delta, \sigma^2_{2g})$ for the 2nd condition, where, following Delmar *et al.* (2005), $|\delta|$ was simulated as a uniformly distributed random variable on the interval (0.25, 0.9). In all simulations the number of over-expressed and under-expressed genes were chosen to be equal. Non differentially expressed genes were simulated under an $N(\mu_g, \sigma^2_{1g})$ distribution for the first condition and $N(\mu_g, \sigma^2_{2g})$ for the second condition. For each simulated data set we calculated p-values by using the default settings of the moderated t-test in the *limma* library. All results are based on 500 simulated data sets.

3.1.2 Sensitivity estimation

As a first step, we simulated a data set with 8 replicates per group and 100 differentially expressed genes. For each of 500 simulations a set of p-values was calculated with the *limma* package and genes were declared significant if their Benjamini-Hochberg (BH) adjusted p-values were below 5%. For this p-value cut-off we then estimated sensitivity and other quantities relating to it according to the methods discussed in Section 2.4.1. Table 1 lists these methods together with names and labels we used for them and the corresponding references.

Table 1: Table of sensitivity estimation methods

Method	Abbr.	Reference
Local false discovery rate, empirical approach	LE	Efron (2004)
Local false discovery rate, theoretical approach	LT	Efron (2004)
Normal mixture, empirical approach	NE	McLachlan *et al.* (2006)
Normal mixture, theoretical approach	NT	McLachlan *et al.* (2006)
Beta Uniform Model	BUM	Pounds and Morris (2003)
Proportion based empirical method	PE	Langaas *et al.* (2005)

Table 2 presents means and standard deviations obtained from the 500 simulations. The first column gives the true observed quantities where "SG" denotes the number of genes found to be significant. The other six columns represent the methods as given in Table 1.

As described before, the number of genes detected was forced to be the same in all methods except when using the Bayesian Uniform Model, so that only that entry varies in the first row of the table. We observe that the number of estimated true positives was fairly similar for all methods, which was also reflected by similar values for the FDR across the different methods. However, the number of estimated false negatives varied considerably, causing similar variation in the sensitivity estimates. We see that, in this simulation, Efron's *locfdr* methods gave the best result. McLachlan's method seemed to over-estimate sensitivity, whereas "PE" and particularly the BUM method underestimated it.

In a second step we varied sample sizes (n = 4, 8, 12) and the number of differentially expressed genes (30, 100, 200), but now only studied the sensitivity estimates. The results in Table 3 show the expected increase in sensitivity as sample size and the number of truly differentially expressed genes go up. Otherwise the results confirmed those of table 2, i.e. again Efron's method gave the best results.

Table 2: Sensitivity estimation from 500 simulated data sets with 8 replicates and 100 differentially expressed genes; the table shows average counts of genes found to be significant (SG), true positives (TP), false negatives (FN), as well as values for the false discovery rate (FDR) and sensitivity (Sens). Standard deviations are given in brackets.

	Truth	LE	LT	NE	NT	BUM	PE
SG	84.7	84.7	84.7	84.7	84.7	78.1	84.7
	(5.3)	(5.3)	(5.3)	(5.3)	(5.3)	(4.8)	(5.3)
TP	79.9	81.6	81.8	79.6	80.3	74.5	81.1
	(4.3)	(4.9)	(4.9)	(4.8)	(4.7)	(4.5)	(5.0)
FN	20.1	26.5	26.9	9.4	11.7	71.2	48.5
	(4.3)	(9.1)	(10.0)	(7.8)	(7.4)	(6.8)	(37.1)
FDR	0.06	0.04	0.03	0.06	0.05	0.05	0.04
	(0.03)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)
Sens	0.80	0.76	0.76	0.90	0.88	0.51	0.67
	(0.04)	(0.06)	(0.07)	(0.06)	(0.07)	(0.02)	(0.15)

These results should be interpreted with caution though, as they might be biased by the way we simulated the micorarray data sets. Robin *et al.* (2007) already pointed out that the empirical *locfdr* sometimes gave unsatisfactory results. We also saw that the order of the sensitivity estimates (with McLachlan's method giving the highest and BUM giving the lowest values) can be quite different when applying the same methods to real data sets. Our main conclusion from this simulation is that the number of false negatives and thus the simulation of sensitivity is quite difficult and that the estimates can vary considerably depending on which method is being used. One possible quality check is to study the FDR estimates. Since our cut-off is based on the Benjamini-Hochberg method, an estimated FDR very different from the predefined threshold is an indication of some problem in the estimation. We found this to be a helpful criterion, particularly when it came to deciding between the theoretical and empirical option in Efron's and McLachlan's approaches.

In subsequent simulations we used Efron's empirical local FDR method for sensitivity estimation since at least for this type of simulated data, it appeared to have good properties.

Table 3: Sensitivity estimation from 500 simulated data sets for varying numbers of replicates (rep) and differentially expressed genes (DE). The table shows average estimated sensitivity (standard deviations).

DE	rep	Truth	LE	LT	NE	NT	BUM	PE
30	4	0.34	0.34	0.23	0.49	0.36	0.10	0.20
		(0.12)	(0.18)	(0.15)	(0.28)	(0.26)	(0.04)	(0.12)
	8	0.74	0.56	0.51	0.82	0.76	0.40	0.46
		(0.09)	(0.15)	(0.16)	(0.18)	(0.21)	(0.5)	(0.20)
	12	0.86	0.64	0.63	0.92	0.89	0.56	0.53
		(0.07)	(0.12)	(0.13)	(0.10)	(0.13)	(0.04)	(0.22)
100	4	0.48	0.52	0.41	0.66	0.58	0.22	0.39
		(0.07)	(0.10)	(0.10)	(0.15)	(0.14)	(0.03)	(0.12)
	8	0.80	0.76	0.76	0.90	0.88	0.51	0.67
		(0.04)	(0.06)	(0.07)	(0.06)	(0.07)	(0.02)	(0.15)
	12	0.89	0.84	0.84	0.95	0.95	0.65	0.74
		(0.03)	(0.06)	(0.06)	(0.03)	(0.04)	(0.02)	(0.14)
200	4	0.57	0.64	0.56	0.77	0.67	0.29	0.51
		(0.04)	(0.05)	(0.07)	(0.09)	(0.08)	(0.02)	(0.09)
	8	0.84	0.86	0.85	0.93	0.91	0.58	0.76
		(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.02)	(0.11)
	12	0.91	0.90	0.90	0.97	0.96	0.70	0.83
		(0.02)	(0.04)	(0.04)	(0.02)	(0.02)	(0.01)	(0.11)

3.1.3 P-value combination methods and simulation of sequential studies

Our next objective was to study how the inverse normal method of combining p-values from different stages performed for simulated data. In a first simulation we compared it against both Fisher's p-value combination method and a *limma* (moderated t-test) analysis that was based on the expression values from all stages (refered to "joint" in tables). We again simulated 3000 genes, of which 100 were over-expressed. There were 6 replicates overall sub-divided into two stages with 3 replicates per group. We emphasise that we did not simulate a stage effect here, i.e. we simulated the optimal situation for the joint *limma* analysis. Table 4 shows the results from 500 data sets.

Table 4: Comparison of 3 different types of meta analysis in simulations of 6 replicates, that were splitted up into 2 stages with 3 replicates per stage. Three thousand genes were simulated 500 times with 100 of them being up-regulated. The table shows average values and standard deviations of counts of significant genes (SG), true positives (TP), false negatives (FN) as well as observed FDR and sensitivity.

	SG	TP	FN	FDR	Sensitivity
Limma	74.95	70.75	29.25	0.06	0.71
	(5.55)	(4.83)	(4.83)	(0.03)	(0.05)
Fisher	83.58	77.11	22.89	0.08	0.77
	(5.68)	(4.57)	(4.57)	(0.03)	(0.05)
Hedges and Olkin	79.78	73.37	26.63	0.08	0.73
	(5.83)	(4.76)	(4.76)	(0.03)	(0.05)

As we can see, the sensitivity is nearly identical between the three methods, so we found no serious loss in power when combining p-values instead of using the full data set. Since Fisher's combination method requires an independent analysis of under-expression and over-expression, we used the inverse normal method in the following. We will refer to it as "meta" in our tables.

In a second simulation we studied how this p-value combination method affects the sensitivity estimation in a sequential set-up. Again 3 replicates were added per group and stage, but a total of 4 stages were studied. We compared the *limma* analysis of the full data set up to each stage with the "meta" method of combining p-values. For both approaches we observed the number of significant genes ("SG") and the true numbers for TP, FDR and sensitivity as well as the corresponding estimated values when using the empirical *locfdr* method. The table shows means (and standard deviations) across 500 simulations. The results (given in Table 5) confirm that the combination of p-values does not cause a serious loss in sensitivity.

The meta method behaved slightly liberal in these simulations, i.e. the observed FDR was higher than the nominal one. The most plausible explanation for this is that the p-values calculated by *limma* are not exact for small sample sizes, so that a method combining these p-values will be more affected than the *limma* analysis which increases sample size at each stage. Still the observed average FDR values remained well below 10%, which we found acceptable. The estimated sensitivity values were below the observed ones for both methods, so our stopping decision behaved conservatively, i.e.

Table 5: Comparison of the meta analysis method with a joint limma analysis in a sequential study with 4 stages and 3 replicates per stage. Three thousand genes with 100 of them being differentially expressed were simulated 500 times. The table shows mean and standard deviation of counts of significant genes (SG), true positives (TP) as well as observed FDR and sensitivity (Sens). For all quantities we give the true value as well as the one estimated by the LE method.

	Stage	Truth joint	LE joint	Truth meta	LE meta
SG	1	23.48(8.61)	23.48(8.61)	23.48(8.61)	23.48(8.61)
	2	75.13(5.86)	75.13(5.86)	79.78(5.76)	79.78(5.76)
	3	88.70(4.81)	88.70(4.81)	94.45(5.00)	94.45(5.00)
	4	94.48(4.35)	94.48(4.35)	100.57(4.95)	100.57(4.95)
TP	1	22.22(7.97)	21.73(7.84)	22.22(7.97)	21.73(7.84)
	2	70.88(5.09)	71.5(5.53)	73.54(4.88)	76.45(5.51)
	3	83.64(3.94)	85.57(4.58)	86.42(3.59)	91.54(4.72)
	4	89.30(3.27)	91.40(3.95)	91.71(3.06)	97.57(4.44)
FDR	1	0.05(0.05)	0.07(0.03)	0.05(0.05)	0.07(0.03)
	2	0.06(0.03)	0.05(0.02)	0.08(0.03)	0.04(0.01)
	3	0.06(0.03)	0.04(0.01)	0.08(0.03)	0.03(0.01)
	4	0.05(0.03)	0.03(0.01)	0.09(0.03)	0.03(0.01)
Sens	1	0.22(0.08)	0.29(0.13)	0.22(0.08)	0.29(0.13)
	2	0.71(0.05)	0.67(0.08)	0.74(0.05)	0.64(0.07)
	3	0.84(0.04)	0.79(0.06)	0.86(0.04)	0.74(0.07)
	4	0.89(0.03)	0.84(0.06)	0.92(0.03)	0.77(0.08)

the experiment was rather continued too long than being stopped prematurely.

For the same simulations we also studied at which stage we would have stopped the experiment according to two different stopping rules. In the first case we stopped if the true or estimated number of true positives exceeded 40, in the second case the criterion was whether the true or estimated sensitivity was at least 60%. The results are summarised in Table 6.

All methods would have stopped at stage 2, when using the first criterion. When using the sensitivity, again most studies would have stopped at stage 2 if the true values had been available. With estimated sensitivity values a small number (2.2%) of all studies would have ended too early, but a larger number of them would have proceeded to stage 3. This again confirmed the

Table 6: For the simulation of table 5 this shows the percentage of studies which have ended at each given stage. The upper part corresponds to a stopping criterion using the estimated number of true positives (TP>40), the lower part uses a sensitivity based criterion (sensitivity > 60%)

Stop. Crit.	Stage	Truth joint	LE joint	Truth meta	LE meta
TP	1	0	0	0	0
	2	100	100	100	100
	3	0	0	0	0
	4	0	0	0	0
Sensitivity	1	0	2.2	0	2.2
	2	97.8	78.4	99.4	68
	3	2.2	19.2	0.6	28.2
	4	0	0.2	0	1.4

conservative nature of the estimation method. Only in very few cases would the full number of stages have been used, which shows the potential of our sequential approach to reduce sample sizes.

3.2 Application to real data sets

Our sequential analysis approach was applied to two real and publicly available data sets, where we artificially split the data into different stages and studied how a sequential analysis would have behaved, if the data had been generated in such a sequential manner. The first experiment was the ApoAI data set, which compares gene expression between apolipoprotein AI (apo AI) knock-out mice and 8 wild type mice from Callow *et al.* (2000). This data set has also been used by Dudoit *et al.* (2002). We used the R package *limma* of Bioconductor to analyse the data, both for the normalisation (print-tip loess normalisation) and the modified t-test and followed the analysis as described in the *limma* tutorial (http://www.bioconductor.org/workshops/2005/labs/lab01/ApoAI.html). We split the data into 3 stages, where the first stage used 4 replicates and stages two and three added two more replicates each.

As in our simulations, we compared the meta-analysis method using the inverse normal combination of p-values with a joint *limma* analysis combining all the expression values up to a given stage. A Benjamini-Hochberg threshold of 0.01 was used in this case. We estimated sensitivity also using the theoretical *locfdr* method for this study (results not given here), but obtained very high

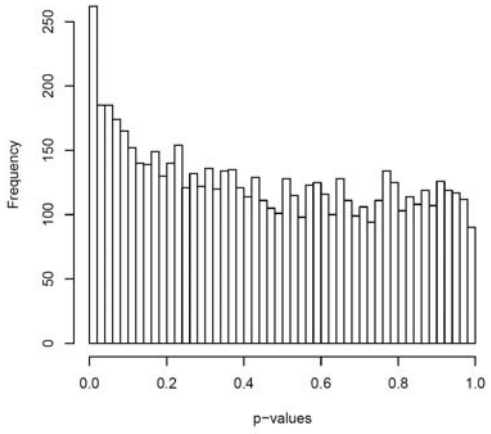
Table 7: Sequential analysis of the ApoAI data set: The table compares a joint limma analysis of the stages with a meta analysis, where sensitivity estimates are based on the LE method. The table gives the numbers of significant genes (SG), true positives (TP), FDR and sensitivity after each stage.

	Stage	Joint	Meta
SG	1	2	2
	2	7	7
	3	8	8
TP	1	1.97	1.97
	2	6.63	6.69
	3	7.69	7.78
FDR	1	0.01	0.01
	2	0.05	0.04
	3	0.04	0.03
Sensitivity	1	0.36	0.36
	2	0.94	0.76
	3	1	1

estimates for the FDR, whereas the empirical *locfdr* approach (cf. table 7) gave reasonable estimates. For this reason our decision criterion used the empirical *locfdr* method.

Both approaches gave a high sensitivity after stage 2 already and in any case we would have stopped the experiment after the 3rd stage, even if it had been possible to add more samples. Here the sets of 2, 7 and 8 significant genes are nested, e.g. the set of 2 genes is a subset of the 7, which again is a subset of the 8 genes. Figure 4 shows a histogram of final p-values for this data set.

Figure 4: Histogram of final p-values for the Apo AI data set after a 3 stage sequential analysis



As a second data set we used the well-known Golub data (Golub *et al.*, 1999). We took the already normalised subset given in the Bioconductor library *multtest* and reduced the data set to 11 patients with acute lymphoblastic leukemia (ALL) and 11 patients with acute myeloid leukemia (AML). We again subdivided the data into three stages with 5, 3 and 3 samples per stage and used a Benjamini-Hochberg threshold of 5%.

When using the empirical *locfdr* method we obtained FDR estimates above 80% after stages 2 and 3. We then investigated the theoretical *locfdr*, which gave estimates below 5%, which were also confirmed by some of the other estimation methods previously discussed as we can see in table 8. This again demonstrates how problematic the estimation of sensitivity can be but also that it can be very useful to check the corresponding FDR estimate as an indicator for such problems.

Figures 5 and 6 illustrate this point very well. The histogram of p-values (Figure 5) has a very distinct peak near zero, clearly indicating a high number of differentially expressed genes. When transforming these p-values to z-scores (Figure 6) they seem to form a unimodal distribution though. The empirical *locfdr* method estimates a null distribution with a large variance in this case (shown on the lower plot in figure 6), that leaves only a very few differentially expressed genes. The theoretical *locfdr* method on the other hand forces the null distribution to be standard normal and thus calls a large number of genes with values in the tails of the overall distribution to be significant (as can be seen in the upper plot of Figure 6).

Table 8: Estimation of sensitivity at each of the three stages of a meta-analysis for the Golub data set. Four different sensitivity estimation methods (LT, NT, BUM, PE) are compared at each stage.

	Stage	LT	NT	BUM	PE
SG	1	13	13	11.77	13
	2	216	216	205.67	216
	3	533	533	525.08	533
TP	1	12.56	12.47	11.48	12.69
	2	208.47	209.6	199.29	209.4
	3	516.49	522.49	511.59	518.5
FDR	1	0.03	0.04	0.02	0.02
	2	0.03	0.03	0.03	0.03
	3	0.03	0.02	0.03	0.03
Sensitivity	1	0.02	0.01	0.01	0.01
	2	0.22	0.17	0.16	0.18
	3	0.42	0.30	0.34	0.38

Figure 5: Histogram of p-values for the Golub data set after a 3 stage sequential analysis

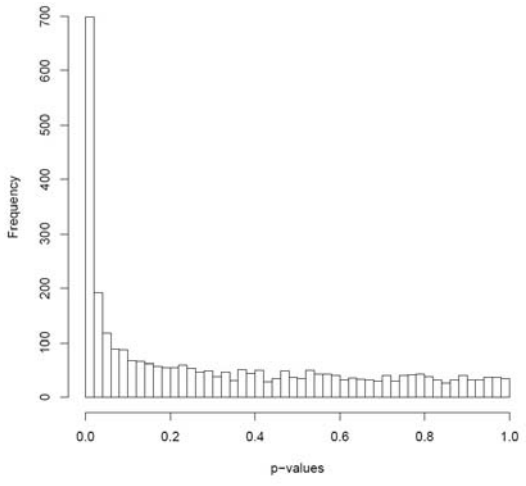
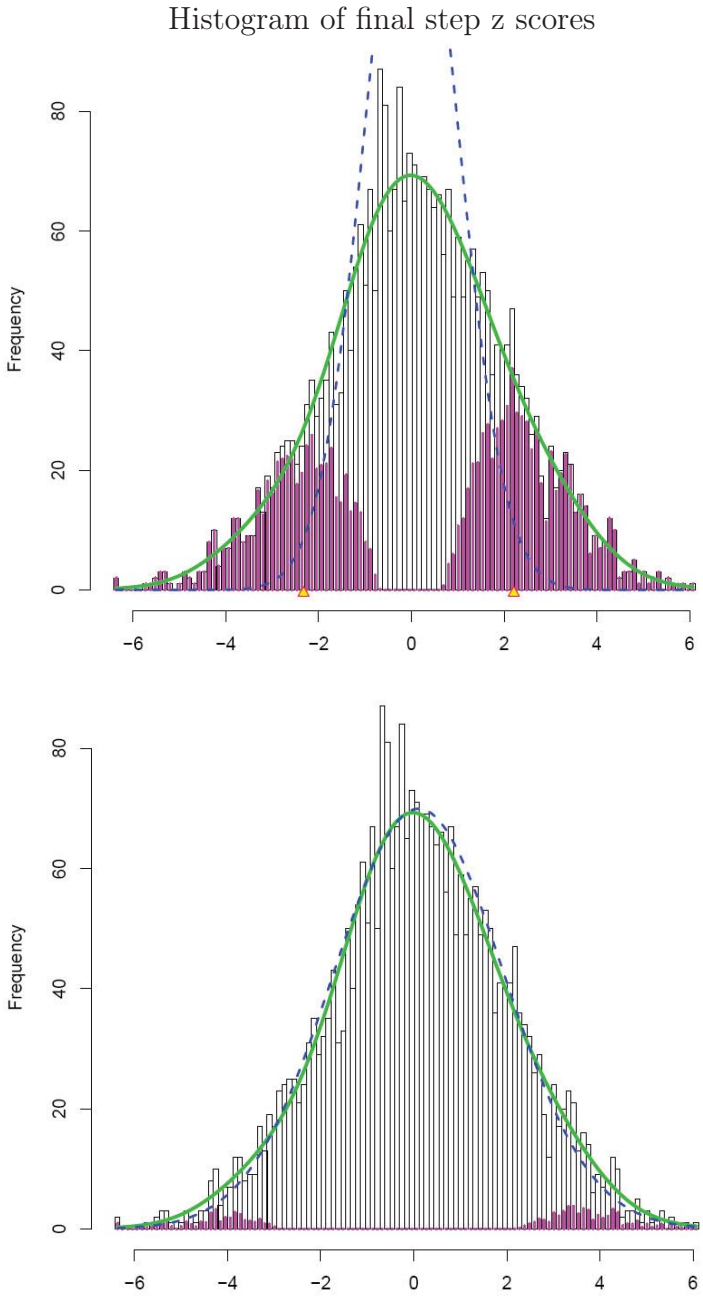


Figure 6: Mixtures fitted to the transformed p-values of the Golub data set using the LT (above) and LE (below) method.



The dashed line represents the standard normal null distribution, the solid line the empirical one.

Since the estimates for the number of genes detected were much higher in this example, the sensitivity estimates were lower, i.e. with a high number of differentially expressed genes it is more difficult to detect them all. In this example it thus would seem more appropriate to use the number of true positives as a stopping criterion since quite a high sample size would be needed to achieve a sensitivity of, say, 80%.

We believe that scientists frequently have an idea whether they are only expecting a very small number of differentially expressed genes like in the Apo-AI data, or whether the two conditions will cause massive changes. But even if this was not the case our sequential approach would give them some indication of the number of differential genes to expect once data from the first stage are available and thus enable them to adapt the strategy and stopping criterion accordingly.

4 Discussion

We suggested a novel strategy that allowed the sequential design and analysis of microarray experiments, an approach that utilised the fact that very often only a limited number of microarrays can be hybridised simultaneously, and allowed us to stop an experiment once sufficient information had been obtained. One key observation is that due to the high number of variables in transcriptomic studies, the stopping decision does not seriously bias the result and thus the main problem of univariate sequential trials does not occur in this context.

Our strategy was based on two main components:

1. At each stage we estimated the sensitivity and the number of true positives as measures of the information obtained up to this stage.

2. We analysed data from different stages by a meta-analysis approach that combined the stage specific p-values

As discussed above, estimation of sensitivity is a difficult problem and different approaches gave varying results. Even though we found that the empirical *locfdr* method gave the best results in our simulations, we saw that it could also produce misleading results, as for example with the Golub data. For this reason we do not recommend a single automatic strategy for sensitivity estimation. This should rather be a supervised process that might take different methods and diagnostic plots into account. One useful criterion in this context is checking the estimation of the FDR. We suggested controlling the FDR by

the Benjamini-Hochberg method, which performs well under relatively general assumptions and makes no assumptions about how p-values of differentially expressed genes are distributed. All sensitivity estimation procedures we considered allowed estimation of the FDR so a comparison of this estimate with the nominal FDR level could help to detect problems in the estimation as we saw in the Golub data example. The estimation of the (local) FDR, sensitivity and other related quantities is a very active area of research and we expect that in future more refined approaches will be available for this part of our analysis strategy.

For the combination of data from different stages, we suggested using the inverse normal method, which had been proposed by Lehmacher and Wassmer (1999) for interim analyses. This p-value combination approach avoids renormalising data from previous stages and the complication of having to model stage effects within the analysis. Our simulations showed that the price that had to be paid for this in terms of power/sensitivity was surprisingly small. We note that this method might also be of interest in a *real* meta-analysis situation, where microarray experiments from genuinely different studies are to be analysed together.

We would like to stress though that the main contribution made by this paper is not the comparison of sensitivity estimation methods or meta-analysis approaches but to present a fairly generic framework that allows a sequential analysis of microarray experiments. We have preferences for options within this framework, but it can be flexibly used. This for example also concerns the type of test being used to detect differential expression. We suggested using the moderated t-test in *limma* but there are a number of alternative approaches that could be used, for example resampling tests or the structural modelling approach in Jaffrézic *et al.* (2007). The only necessary requirement for a test to be used within our strategy, is that it yields valid p-values, i.e. that the p-value distribution under the null hypothesis is uniform.

In this paper we only considered a two-sample problem, but in principle a sequential type of analysis could be used for more complex situations as well. In a situation where more than one effect is being tested (e.g. two different treatments are compared with a control) we will have a p-value distribution and sensitivity estimate for each effect. Based on this, we will have to decide not only whether to add samples but also to which part of the design these samples should be added.

In general, sequential strategies and interim analyses have many potential applications outside the context of clinical trials, for which they were originally developed. In particular our method can be easily adapted to other high throughput technologies.

APPENDIX

We will show why the correlation between an individual p-value and the empirical distribution function of all p-values converges to zero at rate $1/\sqrt{G}$, where G is the number of genes and thus p-values. To illustrate this we assume that the p-values from different genes are independent from each other and we study the empirical distribution function

$$\hat{F}(u) = \frac{1}{G}\sum_{i=1}^{G} 1_{(p_i<=u)}$$

of all p-values for $0 < u < 1$. By the central limit theorem we know that $\sqrt{G}(\hat{F}(u))$ converges to a normal distribution under regularity conditions and so we conclude that $\mathrm{Var}(\sqrt{G}\hat{F}(u))$ converges to a positive constant as G goes to infinity. Now consider an individual p-value (without loss of generality we choose the one for gene 1). Since both p_1 and $(\hat{F}(u))$ are bounded, so is the covariance between the two variables. We thus get for their correlation

$$\sqrt{G}\mathrm{Cor}(p1, \hat{F}(u)) = \frac{\mathrm{Cov}(p_1, \hat{F}(u))}{\sqrt{\mathrm{Var}(\sqrt{G}\hat{F}(u))\mathrm{Var}(p_1)}},$$

which is asymptotically bounded and thus proves the statement.

References

Allison, D. B., Gadbury, G. L., Heo, M., Fernandez, J. R., Lee, C.-K., Prolla, T. A., and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis*, **39**(1), 1–20.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, **57**, 289–300.

Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P., and Rubin, E. M. (2000). Microarray expression profiling identifies genes with altered expression in hdl-deficient mice. *Genome Research*, **10**(12), 2022–2029.

Chang, M. N., Gould, L. A., and Snapinn, S. M. (1995). P-values for group sequential testing. *Biometrika*, **82**(3), 650–654.

Choi, J. K., Yu, U., Kim, S., and Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19 Suppl 1**, 84–90.

Conlon, E. M., Song, J. J., and Liu, A. (2007). Bayesian meta-analysis models for microarray data: a comparative study. *BMC Bioinformatics*, **8**, 80.

Delmar, P., Robin, S., and Daudin, J. J. (2005). Varmixt: efficient variance modelling for the differential analysis of replicated gene expression data. *Bioinformatics*, **21**(4), 502–508.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica Sinica*, **12**, 111–139.

Edgar, R., Domrachev, M., and Lash, A. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, **30**, 207–210.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, **99**(465), 96–104.

Gadbury, G. L., Page, G. P., Edwards, J., Kayo, T., Prolla, T. A., Weindruch, R., Permana, P. A., Mountz, J. D., and Allison, D. B. (2004). Power and sample size estimation in high dimensional biology. *Statistical methods in medical Research*, **13**, 325–338.

Gentleman, R., Ruschhaupt, M., Huber, W., and Lusa, L. (2006). Meta-analysis for microarray experiments. *Bioconductor Vignette*.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., and Lander, E. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**(5439), 531–537.

Hedges, L. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press.

Hess, A. M. and Iyer, H. K. (2007). Fisher's combined p-value for detecting differentially expressed genes using affymetrix expression arrays. *BMC Genomics*, **8**, 96.

Jaffrézic, F., Marot, G., Degrelle, S., Hue, I., and Foulley, J.-L. (2007). A structural mixed model for variances in differential gene expression studies. *Genetical Research*, **89**(1), 19–25.

Lan, K. and DeMets, D. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, **70**, 659–663.

Langaas, M., Lindqvist, B. H., and Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to dna microarray data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(4), 555–572.

Lee, J. W. (1994). Group sequential testing in clinical trials with multivariate observations: a review. *Statistics in Medicine*, **13**(2), 101–111.

Lee, M.-L. T. and Whitmore, G. A. (2002). Power and sample size for dna microarray studies. *Statistics in Medicine*, **21**(23), 3543–3570.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, **55**(4), 1286–1290.

Liu, P. and Gene Hwang, J. T. T. (2007). Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics*, **23**, 739–746.

Loughin, T. M. (2004). A systematic comparison of methods for combining p-values from independent tests. *Computational Statistics and Data Analysis*, **47**(3), 467–485.

McLachlan, G., Peel, D., Basford, K., and Adams, P. (1999). The emmix software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software*, **4**(2).

McLachlan, G. J., Bean, R. W., and Jones, L. B. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, **22**(13), 1608–1615.

Page, G., Edwards, J., Gadbury, G., Yelisetti, P., Wang, J., Trivedi, P., and Allison, D. (2006). The poweratlas: a power and sample size atlas for microarray experimental design and research. *BMC Bioinformatics*, **7**(1), 84.

Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., and Ploner, A. (2005). False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, **21**(13), 3017–3024.

Pounds, S. and Morris, S. W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, **19**(10), 1236–1242.

Robin, S., Barhen, A., Daudin, J., and Pierre, L. (2007). A semi-parametric approach for mixture models: Application to local false discovery rate estimation. *Computational Statistics and Data Analysis*, **51**(12), 5483–5493.

Schäfer, H., Timmesfeld, N., and Müller, H.-H. (2006). An overview of statistical approaches for adaptive designs and design modifications. *Biometrical Journal*, **48**(4), 507–520.

Sébille, V. and Bellissant, E. (2003). Sequential methods and group sequential designs for comparative clinical trials. *Fundamental and Clinical Pharmacology*, **17**(5), 505–516.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**(1).

Spiessens, B., Lesaffre, E., Verbeke, G., Kim, K., and DeMets, D. L. (2000). An overview of group sequential methods in longitudinal clinical trials. *Statistical Methods in Medical Research*, **9**(5), 497–515.

Stevens, J. R. and Doerge, R. W. (2005). Combining affymetrix microarray results. *BMC Bioinformatics*, **6**, 57.

Todd, S. (2007). A 25-year review of sequential methodology in clinical studies. *Statistics in Medicine*, **26**(2), 237–252.

Victor, A. and Hommel, G. (2007). Combining adaptive designs with control of the false discovery rate–a generalized definition for a global p-value. *Biometrical Journal*, **49**(1), 94–106.

Wit, E. and McClure, J. (2004). *Statistics for Microarrays: Design, Analysis and Inference.* Wiley.

Wu, Z., Irizarry, R. A., Gentleman, R., Murillo, F. M., and Spencer, F. (2004). A model based background adjustment for oligonucleotide expression arrays. *Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 1*.

3.4 Complementary results

During the reviewing process of this article, we were asked to perform complementary analyses. In this section, we provide the results which we could not add in the paper due to space constraints.

3.4.1 Normalisation within stages

One of the referees suggested to conduct a within-stage normalisation of the Golub data, but unfortunately the original cel-files for this data set did not seem to be available. In our paper, we used the version of the data provided by the Bioconductor library multtest, which had been pre-processed by the authors of the library to have mean 0 and variance 1 across genes for each array (after log10 and filtering). We thus tried to add a stage effect to the data set by performing a within-stage quantile normalisation on the data instead of scaling each array to have mean 0 and variance 1 (we kept the same other normalisation steps: log10 and filtering). The following two tables give results for an analysis where a) all arrays were quantile normalised simultaneously (see table 3.4) and b) when they were quantile normalised within stages (see table 3.5).

As one can see the differences between the two tables are marginal, but there is quite a difference compared to the non-quantile normalized version presented in our paper. We thus decided not to use this new analysis in the paper as it does not give new insights and we also think that readers might be rather confused if we had used a differently normalised version of a well known data set.

Table 3.4: Sequential analysis of the Golub data after global quantile normalisation

	Stage	Locfdr theo
SG	1	17
	2	200
	3	563
TP	1	16.24
	2	193.38
	3	545.28
FDR	1	0.04
	2	0.03
	3	0.03
Sensitivity	1	0.02
	2	0.19
	3	0.43

Table 3.5: Sequential analysis of the Golub data after within stage quantile normalisation

	Stage	Locfdr theo
SG	1	18
	2	206
	3	568
TP	1	17.18
	2	198.93
	3	549.85
FDR	1	0.05
	2	0.03
	3	0.03
Sensitivity	1	0.02
	2	0.20
	3	0.43

3.4.2 Different design

We were also asked how changing the design would affect our analysis. To answer that question, we repeated one of our analyses for the ApoAI data with different per stage sample sizes (3+3+2 instead of 4+2+2). Corresponding results are given in the following table:

Table 3.6: Sequential analysis of the ApoAI data: the experiment begins with three replicates, three then two replicates are added at the following stages

	Stage	Joint	Meta
SG	1	1	1
	2	7	7
	3	8	8
TP	1	0.99	0.99
	2	6.64	6.75
	3	7.65	7.68
FDR	1	0.01	0.01
	2	0.05	0.04
	3	0.04	0.03
Sensitivity	1	0.45	0.45
	2	0.94	0.79
	3	1	1

Changing the number of replicates in the first stage does not influence the results of the last stage where as many replicates as in the previous design are used. The same 8 genes are detected differentially expressed. The main difference that we observe is the number of differentially expressed genes found at the first stage. The reduction from 4 to 3 replicates leads to one gene less being called significant. Sensitivity found at the first stage in table 3.6 does not seem reliable (it is here overestimated) and generally we would not advise to begin a sequential analysis with only three replicates.

To summarize this part, the sequential analysis of microarray data is to some extent simpler than in the univariate case as the stopping rule does not bias p-values due to the high dimensionality of the data. Thus, results from different stages can be combined by meta-analysis methods and error rates can be controlled by applying standard procedures (e.g. the Benjamini-

Hochberg rule) to the p-values from the combined stages. We suggested stopping rules based on either the estimated number of true positives or the estimated sensitivity. As in clinical trials, sequential designs allow to stop some experiments before the scheduled end and thus save samples. Of course, such an analysis can not be performed when only very few samples are available as it is often the case in animal genetics. It is, however, of great interest in medical studies which can be conducted on more than 10 individuals per condition. Additionally to the cost reduction resulting from a reduced sample size, sequential analysis is also interesting for studies where technical limits cause a staggered availability of the data anyway. This work also provides interesting results in comparing different sensitivity estimation methods which are useful even outside the context of sequential analysis. Although, as our study shows, sensitivity is very difficult to estimate, it is still a useful number to give biologists a rough idea how informative their experiment is. The R package I built for sequential analysis can also be used for sensitivity estimation in a non-sequential design. As we will see in the next chapter, meta-analysis for microarray studies is also an interesting topic itself. Thus this part of the thesis has not only introduced sequential methods to microarray experiments, but also yielded interesting insights into two other areas of general interest: sensitivity estimation and meta-analysis for high-dimensional gene expression data.

Part III

Meta-Analysis

Meta-analysis was the logical follow up of the two previous parts of this PhD. It directly extends sequential analysis since it enables to combine data from different stages. It is also based on shrinkage approaches developped in the first part of my PhD.

Thanks to the high dimensionality of microarray data, the sequential analysis problem was transformed in a need of an appropriate stopping rule and a use of usual meta-analysis methods to combine data from different stages. In the previous paper, data collected at different stages were combined using the inverse normal p-value combination. This p-value combination can also be applied in a more general context to gather data from different studies rather than simple stages. It can combine summary results from studies for which a direct comparison is impossible but which still address the same biological question. The aim of meta-analysis is then double: since the number of replicates is small in most of microarray studies, meta-analysis will increase the sensitivity of the whole experiment including the different studies. What is more, results obtained in the end will be more accurate and common patterns will be drawn. Thus, over the past few years, researchers have tried to combine data from different studies and sometimes across different platforms to gain information. As far as we are concerned, we first concentrate on experiments including studies involving similar platforms and similar chips to avoid problems of annotation which differ between chips. We ask the different studies to answer the same biological question.

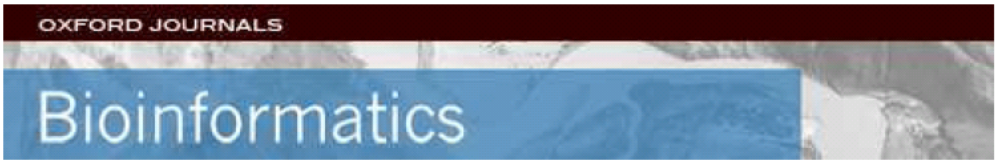
Our methodology is presented in Chapter 4 and Chapter 5 presents an extension to Bayesian models.

Chapter 4

Moderated effect size combination

4.1 Moderated effect size and p-value combinations for microarray meta-analyses

Several approaches are usually considered to combine data. Either data are cross-study normalised and then analysed as a single dataset or summarization results like p-values are combined, using for example Fisher transformation or inverse normal combination detailed in the previous part. An intermediate approach is to model the available expression data including a study effect. The bioconductor package GeneMeta implements one of these modellings (Choi *et al.*, 2003) performing a gene by gene analysis. Since we proved the efficiency of shrinkage approaches in differential gene expression studies (see part I of this PhD), it was natural to try to improve this package by shrinking information from all genes towards common values. The following paper has been submitted to Bioinformatics.



Moderated effect size and p-value combinations for microarray meta-analyses.

BIOINFORMATICS

Vol. 00 no. 00 2009
Pages 1–8

Moderated effect size and p-value combinations for microarray meta-analyses

Guillemette Marot [1*], Jean-Louis Foulley [1], Claus-Dieter Mayer [2] and Florence Jaffrézic [1]

[1]INRA, Génétique Animale et Biologie Intégrative, Jouy-en-Josas, F-78350, France.
[2]Biomathematics and Statistics Scotland, UK.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: With the proliferation of microarray experiments and their availability in the public domain, the use of meta-analysis methods to combine results from different studies increases. In microarray experiments, where the sample size is often limited, meta-analysis offers the possibility to considerably increase the statistical power and give more accurate results.

Results: A moderated effect size combination method was proposed and compared to other meta-analysis approaches. All methods were applied to real publicly available datasets on prostate cancer, and were compared in an extensive simulation study for various amounts of inter-study variability. Although the proposed moderated effect size combination improved already existing effect size approaches, the p-value combination was found to provide a better sensitivity and a better gene ranking than the other meta-analysis methods, while effect size methods were more conservative.

Availability: An R package metaMA is available on the CRAN.

Contact: guillemette.marot@jouy.inra.fr

1 INTRODUCTION

Meta-analysis, which consists in combining data or results from different studies, has been widely used in medicine and health policy to interpret contradictory results from various studies or overcome the problem of reduced statistical power in studies with small sample sizes. Hedges and Olkin (1985) and Stangl and Berry (2000) provide good reviews of meta-analysis techniques.

Microarray experiments are a typical example of small sample size designs. These experiments, which enable us to study gene expressions from thousands of genes at a time, often rely on very few samples due to their high cost or the lack of biological replicates available. Choi *et al.* (2003) and Rhodes *et al.* (2002) were among the first authors to raise the issue of meta-analysis in the context of microarray data to find differentially expressed genes. Some of Choi *et al.* (2003) methods are implemented in the Bioconductor package GeneMeta described by Lusa *et al.* (2008). These approaches rely either on combinations of expression values themselves and the modelling of the inter-study variation using effect size calculations (Choi *et al.*, 2003) or on the combination of p-values (Rhodes *et al.*, 2002). Conlon *et al.* (2007), Scharpf *et al.* (2007) also proposed Bayesian methods to combine microarray data.

In the last few years, several authors such as Smyth (2004) or Jaffrézic *et al.* (2007) showed that, in single study analyses, shrinkage approaches leading to moderated t-tests were more powerful than gene-by-gene methods to detect differentially expressed genes when small numbers of biological replicates are available. Indeed, shrinkage consists in estimating each individual gene value by taking into account information from all genes of the experiment. By decreasing the total number of parameters to estimate, this increases sensitivity, that is to say the proportion of true positives among the truly differentially expressed genes. In the previously mentioned meta-analyses studies, authors based the calculation of the p-values or effect sizes to be combined on standard t-tests, i.e. on gene-by-gene analyses. They therefore gained sensitivity for gene detection by combining different studies but it is expected that even more sensitivity could be obtained using shrinkage approaches. The aim of this paper is to propose a method to calculate moderated effect sizes and to compare their performance with the combination of standard effect sizes or of p-values from standard and moderated t-tests. These methods were applied to publicly available datasets on prostate cancer and compared in an extensive simulation study.

2 METHODS

2.1 Effect size calculation

Let Y_{sigr} and Y_{sjgr} be the expression levels for gene g in conditions i and j for study s and replicate r. The data are assumed to be normally distributed as $Y_{sigr} \sim \mathcal{N}(\mu_{sig}, \sigma^2_{sg})$ and $Y_{sjgr} \sim \mathcal{N}(\mu_{sjg}, \sigma^2_{sg})$. A simple effect size is the standardized difference:

$$\delta_{sg} = (\mu_{sig} - \mu_{sjg})/\sigma_{sg} \tag{1}$$

For effect size calculations, the procedure described by Choi *et al.* (2003) was applied to estimate the study effect and obtain a test statistic for differential expression. The corresponding hierarchical model used was therefore:

$$\begin{aligned} d_{sg} &= \theta_{sg} + e_{sg}, \quad e_{sg} \sim \mathcal{N}(0, w^2_{sg}) \\ \theta_{sg} &= \mu_g + v_{sg}, \quad v_{sg} \sim \mathcal{N}(0, \tau^2_g) \end{aligned} \tag{2}$$

*to whom correspondence should be addressed

where d_{sg} is the estimation of the effect size for study s and gene g, τ_g^2 represents the variability between studies while w_{sg}^2 are the within-study variances. These within-study variances have already been estimated in the same stage as the estimation of the effect size. They are therefore assumed to be known in the hierarchical model, which makes the difference with a linear mixed model. An estimation of the between-study variances τ_g^2 can be obtained using the method of moments as suggested by Choi *et al.* (2003). Parameter μ_g is estimated as in the generalized least squares method: $\widehat{\mu_g}(\tau_g^2) = \sum(w_{sg}^2 + \tau_g^2)^{-1} d_{sg} / \sum(w_{sg}^2 + \tau_g^2)^{-1}$, with $Var(\widehat{\mu_g}(\tau_g^2)) = 1/\sum(w_{sg}^2 + \tau_g^2)^{-1}$. A z-score to test for differentially expressed genes is then constructed as follows:

$$z_g = \frac{\widehat{\mu_g}(\tau_g^2)}{\sqrt{Var(\widehat{\mu_g}(\tau_g^2))}} \tag{3}$$

Although Choi *et al.* (2003) advise permutations to calculate p-values and estimate the FDR, a faster solution is suggested in the Bioconductor package GeneMeta, which assumes a normal distribution on the z-scores after checking the reliability of this hypothesis by a q-q plot.

Moderated effect sizes for unpaired data To estimate the effect size defined in equation (1) for unpaired data, Choi *et al.* (2003) considered the unbiased estimator of the standardized mean difference (for more clarity, indices g and s are omitted in this section) :

$$d' = d(1 - 3/(4(n-2)-1)) \tag{4}$$

where $d = (\bar{Y}_i - \bar{Y}_j)/S_p$ for conditions i and j, and S_p are pooled standard deviations. These d effect sizes can easily be linked to Student t-tests via the relationship:

$$t = d\sqrt{\tilde{n}} \tag{5}$$

with $\tilde{n} = n_i n_j/(n_i + n_j)$ where n_i (resp. n_j) is the number of replicates in condition i (resp. j).

We propose to extend these effect sizes to account for moderated t-tests. We first consider the popular shrinkage approach proposed by Smyth (2004) and implemented in the Bioconductor R package limma. We will also accomodate the effect size calculation to another shrinkage approach proposed by Jaffrézic *et al.* (2007), which allows us to analyze data with heterogenenous variances between conditions. This method is implemented in the R package SMVar available on the CRAN.

As the same variance is assumed for both conditions in limma, in this case the moderated effect size can be estimated as:

$$d_{Limma} = t_{Limma}/\sqrt{\tilde{n}} \tag{6}$$

For SMVar, different variances are assumed in each condition i and j such that: $Y_{ir} \sim \mathcal{N}(\mu_i, \sigma_i^2)$ and $Y_{jr} \sim \mathcal{N}(\mu_j, \sigma_j^2)$. In this case, we rely on the effect size definition proposed by Kulinskaya and Staudte (2007) where the denominator σ is:

$$\sigma = \{\frac{q\sigma_i^2 + (1-q)\sigma_j^2}{q(1-q)}\}^{1/2} \tag{7}$$

with $q = n_j/n$ and $n = n_i + n_j$. This parameter can be rewritten as $\sigma^2 = (n_j\sigma_i^2 + n_i\sigma_j^2)/\frac{n_i n_j}{n}$ so that $\frac{\sigma^2}{n} = \frac{\sigma_i^2}{n_i} + \frac{\sigma_j^2}{n_j}$. This effect size can therefore be linked to the Welch statistic as:

$$t_{Welch} = \sqrt{n}\, d_{Kulinskaya} \tag{8}$$

As SMVar relies on a Welch statistic, a natural moderated effect size would be:

$$d_{SMVar} = t_{SMVar}/\sqrt{n} \tag{9}$$

To apply the meta-analysis procedure described in the previous paragraph, variances of effect sizes are also needed. The estimator of the variance $Var(d) = (n_i^{-1} + n_j^{-1}) + d^2(2(n_i + n_j))^{-1}$ given in Choi *et al.* (2003) is, however, an asymptotic estimator. As the number of replicates is often limited in microarray experiments, we decided to compute the exact form of the variances for moderated effect sizes. Using the distribution of effect sizes provided by Hedges (1981), it can be shown that:

$$Var(d) = \frac{m}{(m-2)\tilde{n}}[1 + \tilde{n}\delta^2] - \delta^2/[c(m)]^2 \tag{10}$$

with

$$c(m) = \frac{\Gamma(\frac{m}{2})}{\sqrt{\frac{m}{2}}\Gamma(\frac{m-1}{2})}. \tag{11}$$

In these formulae, δ is the effect size defined in equation (1), $\tilde{n}$ is equal to $n_i n_j/(n_i + n_j)$ for limma and to $n = n_i + n_j$ for SMVar, and m is the number of degrees of freedom. Note that for both estimators given in equations (6) and (9), the calculation of this variance is possible using equation (10) thanks to the degrees of freedom of the moderated t-statistics provided in both procedures. For limma (Smyth, 2004), m equals to the sum of prior degrees of freedom and residual degrees of freedom for the linear model for gene g. For SMVar, degrees of freedom are calculated by Satterthwaite's approach as:

$$m = \frac{(\frac{\widehat{\sigma_i}^2}{n_i} + \frac{\widehat{\sigma_j}^2}{n_j})^2}{\frac{\widehat{\sigma_i}^4}{n_i^2}\frac{V(\mathbf{ln}\widehat{\sigma_i}^2)}{2} + \frac{\widehat{\sigma_j}^4}{n_j^2}\frac{V(\mathbf{ln}\widehat{\sigma_j}^2)}{2}}. \tag{12}$$

This generalizes the formula given in Jaffrézic *et al.* (2007) for the case where the number of replicates is the same for both conditions.

Unbiased estimators of effect sizes Using the distribution of effect sizes provided by Hedges (1981), unbiased estimators can be defined from the previously proposed moderated effect sizes as:

$$d'_{moderated} = c(m) d_{moderated}. \tag{13}$$

with $c(m)$ given in equation (11). Equation (13) can be seen as an extension of equation (4) with $d' = c(m)d$ where $c(m) = 1 - 3/(4m-1)$ and $m = n - 2$.
Assuming that $Var(c(m)) = 0$, which holds exactly for standard effect sizes and works quite well in practice for moderated effect sizes, the variance of the unbiased effect sizes is computed as $c(m)^2$ times the variance of the biased estimators given in equation (10). Since $c(m) < 1$, unbiased estimators have a smaller variance than biased ones.

Moderated effect sizes for paired data For both moderated t-tests with limma and SMVar, the unbiased effect size for paired data is obtained via the relationship:

$$d'_{paired} = c(m)\frac{t_{moderated}}{\sqrt{n_{paired}}} \tag{14}$$

with n_{paired} the number of replicates.

2.2 p-value combination

Many authors such as Rhodes *et al.* (2002) and Hu *et al.* (2006) use Fisher's combined probability test to combine p-values across studies. The main disadvantage of this approach is that, as pointed out by Hong and Breitling (2008), it requires to treat over and under-expressed genes separately. We therefore suggest to use the inverse normal method in this study, which is symmetric in the sense that p-values near zero are accumulated in the same way as p-values near unity (Hedges and Olkin, 1985), and is therefore suitable for combining results for differentially expressed genes when the direction of deviation from the null hypothesis is not known. Loughin (2004) compared different p-value combination methods in a simulation study and stated that the (unweighted) inverse normal method worked well in cases where the evidence against the null-hypothesis is spread equally across the different studies. The inverse normal method, so-called by Hedges and Olkin (1985), refers to the averaging of transformed individual p-values to normal scores. This procedure was first introduced independently by Stouffer *et al.* (1949) and by Liptak (1958). Let N_s be the total number of studies to be combined and $\tilde{p}_g(s)$ the individual p-value calculated for study s and gene g.

$$S_g = \frac{1}{\sqrt{N_s}} \sum_{s=1}^{N_s} \Phi^{-1}(\tilde{p}_g(s)) \tag{15}$$

To avoid directional conflicts, it is necessary to use one-sided p-values for each study. Under the null hypothesis S_g follows a standard normal distribution. An overall two-sided p-value can then be obtained by

$$p_g = 2(1 - \Phi(|S_g|)) \tag{16}$$

An alternative to (15) is to use the weighted method by Marot and Mayer (2009), which is implemented in the R package metaMA.

$$S_g = \sum_{s=1}^{N_s} w_s \Phi^{-1}(1 - \tilde{p}_g(s)) \tag{17}$$

with

$$w_s = \sqrt{\frac{n(s)}{\sum_{i=1}^{N_s} n(i)}}$$

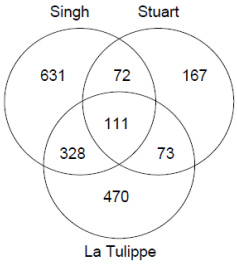
where $n(s)$ is the number of replicates in study s.

3 APPLICATION TO REAL DATASETS

These different methods were compared on real data sets on prostate cancer. Datasets from Singh *et al.* (2002), LaTulippe *et al.* (2002), Stuart *et al.* (2004) were downloaded from public websites. All these experiments were generated from the same Affymetrix HG_U95Av2 platform. Data from CEL files were normalized using RMA (Irizarry *et al.*, 2003). In the following, datasets are referred to by the name of the corresponding first author. The Singh dataset contains 102 samples, 50 of which are non tumor prostate samples, the other 52 being prostate tumors. LaTulippe provides 3 normal samples and 23 cancer samples while there are 50 normal and 38 cancer samples in the Stuart dataset. Only the 12600 genes in common between the three datasets were kept for the analysis. Although these datasets are not representative of small sample size designs, we used them to extract real inter-study variation and to illustrate how the methods proposed here can be applied. Simulations with smaller sample size designs are presented in the next section.

We first performed standard limma analyses for each of the three studies and applied a Benjamini Hochberg (BH) correction to take into account the multiple testing problem. At a 1% Benjamini-Hochberg (BH) threshold, 1852 genes were significant, 1142 of which in the Singh study, 423 and 982 in the LaTulippe and Stuart studies, respectively. As shown in the venn diagramm given in figure 1, only 111 genes were found in common between these three datasets.

Fig. 1. Venn diagram comparing the lists of differentially expressed genes at a 1% BH threshold obtained by each individual study



When binding all the expression data together and including a study effect in the limma linear model, 2422 genes were found significant at the same BH threshold. We compared this gene list with the ones obtained with 1) effect size combination; 2) weighted inverse normal p-value combination, both procedures being based on limma moderated t-tests. For the effect size combination, 1487 genes were found to be differentially expressed at a 1% BH threshold and for the p-value combination, 2637 genes were found significant. Venn diagram corresponding to the comparison of these methods is given in Figure 2. It was found that 1427 genes were common between the three approaches. It can also be noticed that the p-value combination method detected all the genes found with the effect size combination method and all but 101 with the limma including study effect analysis. On the other hand, 256 genes were detected only by the p-value combination approach.

Fig. 2. Venn diagram comparing the lists of differentially expressed genes at a 1% BH threshold obtained by combining p-values, effect sizes or binding expression data together

More results obtained with the standard and the proposed moderated effect sizes, p-value combination or a global analysis using standard and moderated t-tests are given in Table 1. Since many replicates were involved, we could not observe on these real datasets the gain of differentially expressed genes usually found with shrinkage approaches. We could check that, in this case, using the exact variance for standard effect sizes did not change much the number of differentially expressed genes compared to using the asymptotic variance. Indeed, the proposed effect size combination based on usual t-tests and the exact variance detected 1507 differentially expressed genes while the z-score given by the GeneMeta package found 1498 differentially expressed genes. Table 1 also points out that p-value combination methods detected more genes than either expression data combination or effect size combination.

Table 1. Number of differentially expressed genes for the real dataset provided by the different meta-analysis approaches at a 1% BH threshold

	effect size combination	p-value combination
T-test	1507	2623
Limma	1487	2637
SMVar	1647	2730

As far as gene rankings were concerned, they were very similar. Spearman rank correlations equalled 0.81 between the expression data and effect size combination absolute values of test statistics and equalled 0.92 and 0.91 between the p-value combination and each of the two other methods, respectively. Only 33 genes differed between limma including study effect and p-value combination top 1000 gene lists. Effect size combination appeared to have a slightly different ranking with 353 out of its top 1000 genes not found in the other top gene lists. Between the three methods, 554 differentially expressed genes were found in common in the top 1000 gene lists.

4 SIMULATIONS

Expression data were simulated using a hierarchical model:

$$\begin{aligned} y_{sigr} &= \Theta_{sig} + \epsilon_{sigr}, \quad \epsilon_{sigr} \sim \mathcal{N}(0, Vintra_g) \\ \Theta_{sig} &= \mu_{ig} + u_{sig}, \quad u_{sig} \sim \mathcal{N}(0, Vinter_g) \end{aligned} \tag{18}$$

where y_{sigr} is the expression level for replicate r of gene g in condition i and study s. Note that the variances specified in (2) are related but not identical to (18) because the equations in (2) model effect sizes while the equations in (18) model expression values. Parameters of simulation were obtained from the three real datasets analyzed in the previous section. For μ_{ig}, we considered the empirical means of gene expression values observed in each condition (tumor/normal) of these datasets. Mean expression values were supposed to be equal for all genes but the 1427 genes previously found in common between the limma including study effect analysis, the effect size and p-value combination methods (Figure 2). For the genes simulated as non-differentially expressed, means in both conditions were equal to the average of the empirical means of the two conditions of the Singh dataset. Variance parameters were calculated from the three real datasets and kept different for each gene. The within-study variances were equal to the gene-by-gene empirical estimations of variances in these datasets and were kept different per gene, condition and study. Between-study variance was simulated as the observed between-study variance averaged over the two conditions.

Expression data combination, standard and moderated effect size combination, as well as p-value combination based both on standard and moderated t-tests were compared in a simulation study with 300 runs. For each method, the number of True Positives (TP), False Positives (FP), False Negatives (FN) and Sensitivity were calculated. All these criteria were defined as in Marot and Mayer (2009). In particular sensitivity was defined as follows:

$$\text{Sensitivity} = E(\frac{TP}{FN+TP}) \tag{19}$$

As a compromise between False Discovery Rate and Sensitivity, we also calculated the area under the ROC curve (AUC) for each method. A few plots of ROC curves are given later in the paper. To draw ROC curves, the number of False Positives, True Positives, False Negatives and True Negatives were computed for all possible cut-offs in the gene list (1-5000). This procedure was repeated for the 300 simulations and the curves describe the dependency between sensitivity $E(\frac{TP}{TP+FN})$ and specificity $E(\frac{TN}{TN+FP})$. The higher the area under the curve is, the better the gene ranking is.

In these simulations we considered 3 or 5 studies and 6, 8 or 10 replicates for each study. When five studies were simulated, parameters from the third and the fourth study equalled the ones extracted from the Singh and LaTulippe datasets, respectively. For simplicity, the same number of replicates was simulated in each condition for all studies.

Table 2. Influence of the number of studies and replicates (Rep) on the comparison of meta-analysis methods using moderated effect size (ES) estimators for a BH threshold of 5%. The table shows average estimated sensitivity (Sens), FDR and area under ROC curve (AUC) as well as their estimated standard deviations into brackets on 300 simulations.

	Rep	(%)	ES	ES_{Limma}	ES_{SMVar}
3 studies	6	Sens	1.1(0.5)	4.8(0.9)	7.7(1)
		FDR	0.2(1)	0.8(1.1)	1.3(1)
		AUC	82.7(0.6)	83.3(0.6)	83.4(0.6)
	8	Sens	7.0(1.1)	11.0(1.2)	13.8(1.3)
		FDR	0.9(0.9)	1.2(0.9)	1.5(0.9)
		AUC	86.4(0.5)	86.8(0.5)	86.9(0.5)
	10	Sens	14.2(1.3)	17.5(1.3)	20.0(1.4)
		FDR	1.4(0.8)	1.6(0.8)	1.8(0.8)
		AUC	89.0(0.5)	89.3(0.5)	89.4(0.5)
5 studies	6	Sens	14.3(1.2)	22.9(1.4)	26.7(1.4)
		FDR	0.5(0.4)	1.0(0.5)	1.5(0.7)
		AUC	91.2(0.5)	91.6(0.4)	91.6(0.4)
	10	Sens	47.2(1.4)	50.4(1.3)	52.1(1.3)
		FDR	1.6(0.5)	1.7(0.5)	1.9(0.5)
		AUC	95.8(0.3)	95.9(0.3)	95.9(0.3)

It can be seen from Table 2 that the fewer replicates there were, the larger the gain in sensitivity due to the shrinkage of effect sizes was. Indeed, for five studies and six replicates in each condition, the average sensitivity when using classical effect sizes was 14.3% and it increased to 22.9% and 26.7% when shrinking with limma and SMVar, respectively. All false discovery rates were below the 5% threshold that was required, slightly higher for the SMVar approach and lower for the standard effect size method. Similar ranking of the methods was observed when changing from five to three studies, with an increase in sensitivity for all methods for a given number of replicates. For example, the sensitivity for the standard effect size approach with ten replicates per condition was 47.2% for five studies while only 14.2% when based on three studies. This was expected since the total number of replicates used in the meta-analysis was smaller in the latter case. Gene ranking was very slightly improved by using moderated effect sizes. For example, when only six replicates and three studies were considered, the AUC calculated after a gene-by-gene meta-analysis equalled to 0.827 and increased to 0.834 with a moderated effect size combination.

In the second set of simulations, we studied the influence of between-study variability, comparing all the methods either on datasets where no-inter study was simulated or on datasets simulated as previously, accounting for between-study variation. When simulating an homogeneous dataset, among the previous 1427 genes, only the genes significant at a 1% BH threshold in the Singh dataset were simulated differentially expressed.

As shown in Table 3, presence of inter-study variability especially influenced the performance of data expression combination analyses. In these simulations, sophisticated meta-analysis methods were compared to naive methods denoted "joint" which gathered all the expression data together whithout taking into account any study effect. Since the limma package allows to include a study effect in the linear model, we defined two types of limma joint analyses. $Joint_{L1}$ referred to the very naive approach, while $Joint_{L2}$ was the limma global analysis including the study effect in the linear model. This second approach can be viewed as an alternative meta-analysis method. As expected, when there was no inter-study variability, it was better to bind the expression data from all studies and perform a joint analysis, whatever the method chosen, to find differentially expressed genes than to combine effect sizes. Indeed, the area under the curve was higher for the joint analyses (0.994) than for the effect size combination methods. Since effect size combinations are more conservative, a larger difference was observed in terms of sensitivity: it was around 84% for joint analyses, whereas it was only around 70% for effect size methods. When simulating between-study variability with parameters extracted from the real prostate cancer datasets, the situation was reversed and meta-analysis methods gave better sensitivities and better AUC than joint analyses. The p-value combination methods outperformed effect size combination methods with a sensitivity around 69 to 72% for the first ones and 47 to 52% for the latter ones. False discovery rates were a bit higher for p-value combination methods, which reflects the fact that the expression value and effect size combinations were much more conservative than the p-value combination. These False Discovery Rates were, however, still around the required 5% Benjamini-Hochberg threshold. We also noticed that shinkage improved the meta-analysis approaches within the same scheme (either p-value or effect size combination). The AUC results confirmed the good performance of p-value combination methods, showing that the gain in sensitivity is not uniquely an artefact of a higher False Discovery Rate.

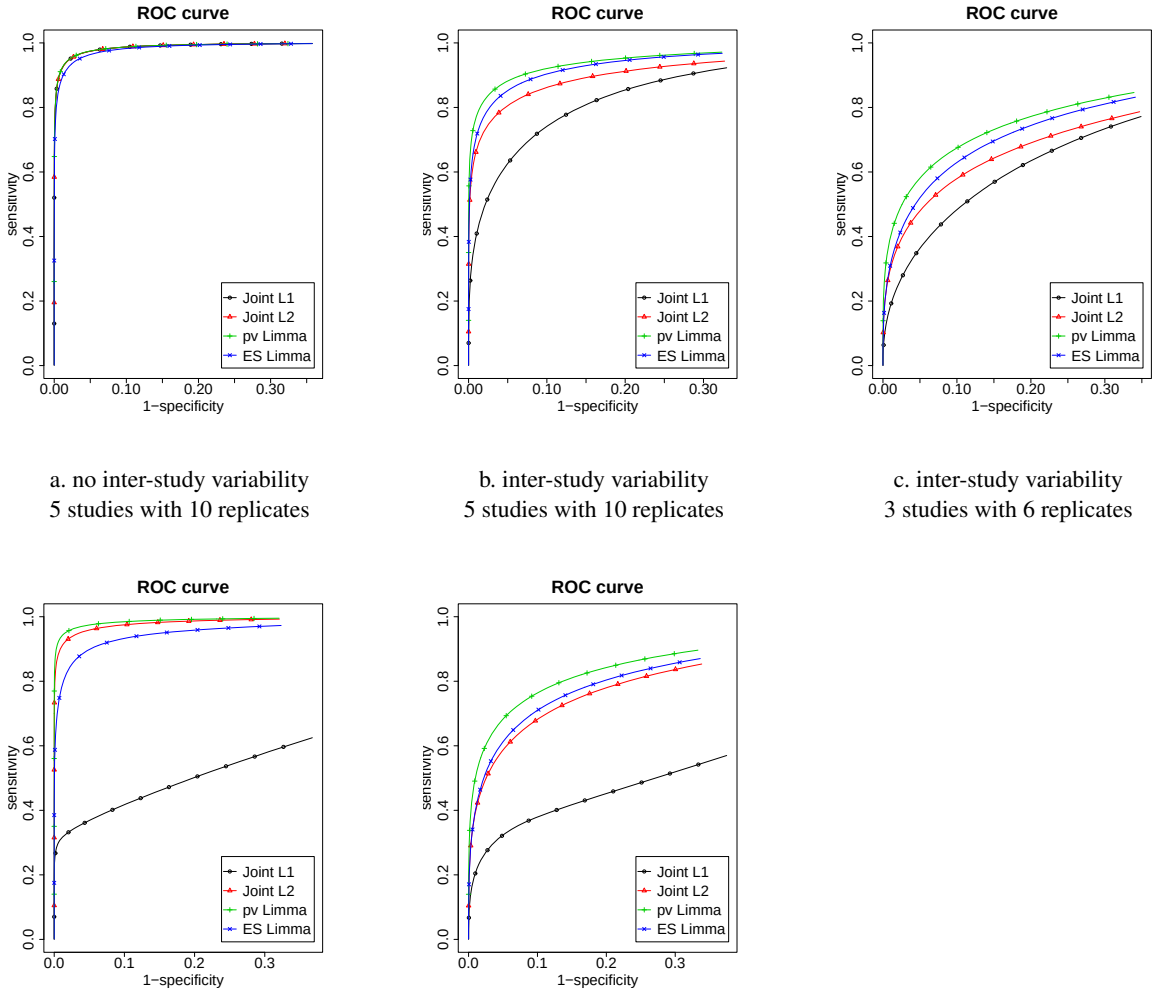
Table 3. Influence of the presence of between-study variability (inter) on 300 simulations with 10 replicates for both conditions in each of the 5 studies. $Joint$ and $Joint_{SMVar}$ denote the t-test and the SMVar global analyses, respectively. $Joint_{L1}$ and $Joint_{L2}$ are the global limma analyses, the first one only gathering the expression data, the second one including a study effect in the linear model. ES stands for effect size combination and pv for p-value combination.

inter		Sens (%)	FDR (%)	AUC (%)
no	$Joint$	83.7(1.4)	4.7(0.9)	99.4(0.1)
	$Joint_{SMVar}$	84.8(1.4)	4.7(0.9)	99.4(0.1)
	$Joint_{L1}$	84.2(1.5)	4.7(0.9)	99.4(0.1)
	$Joint_{L2}$	84.1(1.4)	4.7(0.9)	99.4(0.1)
	ES	66.8(1.9)	1.5(0.5)	99.2(0.2)
	ES_{SMVar}	71.6(1.8)	2.2(0.7)	99.2(0.2)
	ES_{Limma}	69.8(1.8)	1.9(0.6)	99.2(0.2)
	pv	82.8(1.5)	4.8(0.9)	99.3(0.1)
	pv_{SMVar}	87.1(1.3)	6.1(0.9)	99.5(0.1)
	pv_{Limma}	84.8(1.4)	4.9(0.9)	99.4(0.1)
yes	$Joint$	3.9(0.6)	0.1(0.4)	89.8(0.4)
	$Joint_{SMVar}$	3.9(0.7)	0.0(0.3)	90.0(0.4)
	$Joint_{L1}$	3.8(0.7)	0.0(0.3)	90.0(0.4)
	$Joint_{L2}$	57.2(1.2)	4.3(0.7)	93.9(0.4)
	ES	47.2(1.4)	1.6(0.5)	95.8(0.3)
	ES_{SMVar}	52.1(1.3)	1.9(0.5)	95.9(0.3)
	ES_{Limma}	50.4(1.3)	1.7(0.5)	95.9(0.3)
	pv	69.3(1.1)	4.7(0.7)	96.4(0.3)
	pv_{SMVar}	72.9(1)	5.3(0.7)	96.6(0.3)
	pv_{Limma}	71.2(1)	4.6(0.6)	96.6(0.3)

AUC numbers were illustrated by the ROC curves plotted in Figure 3 (see a. and b.), which show the performance in terms of gene ranking. We only focused on meta-analysis methods based on the limma moderated t-test, since limma is the most commonly used package for differential expression. While all curves were similar when no inter-study variability was simulated, p-value combination slightly outperformed the other approaches as soon as between-study variability was introduced.

Similar results on the relative performance of meta-analysis methods were obtained with fewer replicates or different numbers of replicates per studies and conditions. For example, we simulated 3 studies with 6 replicates in each (see Figure 3 c.). We also tested another design with 10 replicates in both conditions of study 1, 10 (resp.8) replicates in condition 1 (resp.2) of study 2 and 3 and 9 replicates in study 3 (see Figure 3 e.). With these settings, effect size combination also outperformed a joint limma analysis including a study effect, which was already observed in Table 3. It has to be pointed out that this simulation study tends to favour the last method as a simple additive study effect was considered. The p-value and effect size combination methods might perform much better than the simple limma linear study effect model in more complicated settings. Nevertheless, we found particularly interesting the fact that when adopting the real application design, that is to say 50 and 52 replicates in study 1, 3 and 23 in study 2 and 50 and 38 in study 3,

Fig. 3. ROC curves comparing gene ranking with various settings for number of replicates and inter-study variability:



a. no inter-study variability
5 studies with 10 replicates

b. inter-study variability
5 studies with 10 replicates

c. inter-study variability
3 studies with 6 replicates

d. inter-study variability
real application design

e. inter-study variability
10/10 replicates (conditions 1/2) in study 1,
10/8 and 3/9 in studies 2 and 3

the position of the curves was reversed between the last two methods (see Figure 3 d.). In all cases, the p-value combination provided a better gene ranking than the other combination approaches.

To evaluate the performance of meta-analysis methods, Choi *et al.* (2003) and Conlon *et al.* (2007) defined the integration-driven discovery rate (IDR) as the proportion of genes that are identified as differentially expressed (DE) in the meta-analysis that were not identified in any of the individual studies alone. In the same way, Stevens and Doerge (2005) and Conlon *et al.* (2007) defined the integration-driven revision rate (IRR) as the percentage of genes that are declared DE in individual studies but not in meta-analysis. While IDR represents the information gained by meta-analysis, IRR measures the loss due to it.

$$\text{IDR} = \frac{\#\text{genes[DE in MA and non DE in any IS]}}{\#genes[DE\ in\ MA]} \tag{20}$$

$$\text{IRR} = \frac{\#\text{genes[DE in at least one IS and non DE in MA]}}{\#\text{genes[DE in at least one IS]}} \tag{21}$$

In these formulae MA refers to Meta-Analysis and IS to Individual Studies. We found that interpretating Integration Discovery Rates was quite misleading since they are highly dependent on the number of differentially expressed genes found with each method. Discoveries or Revisions, which correspond to the numerators of the previous quantities are therefore given here in addition to these rates.

In this last part of the simulation study, we considered five studies, with 10 replicates per study and between-study variability close to the one observed in the real application. Results presented in Table 4 show that, in this setting, individual studies missed out many genes. Even if all studies had the same number of replicates, large differences could still be observed between the different studies, depending on the within-study variances adopted for each study. In particular, the first and the fourth studies, whose within-study variances had been simulated from the Singh dataset, detected very few genes compared to the other ones, with 16.5 genes DE on average over the 300 simulations, corresponding to a sensitivity of 1.1%. On the other hand, the second and fourth studies, whose within-study variances had been simulated from the LaTulippe dataset, had the highest sensitivity, equal to 18.4%. From the last column "summary" of the table it can be noticed that most of the genes found in the individual analyses were different from one study to the other since there was an average of 469.6 genes found in total when pooling the individual lists. In this global list, the False Discovery Rate was higher than the 5% BH required threshold, which was expected as there was no further correction after combining gene lists.

Table 4. Results with limma analyses for individual studies (10 replicates for both conditions in each study). Column "summary" shows the number of genes obtained when pooling the lists of DE genes from individual studies.

	Study 1	Study 2	Study 3	Study 4	Study 5	Summary
DE	16.5(8.9)	273.7(16.7)	162.3(9.9)	16.0(8.7)	273.6(17.2)	469.6(18.8)
Sens.	1.1(0.6)	18.4(1.1)	10.7(0.6)	1.1(0.6)	18.4(1.1)	30.5(1.1)
FDR	4.4(5.4)	4.2(1.2)	6.1(2)	4(5.3)	4.2(1.2)	7.4(1.1)

On the other hand, Table 5 shows that performing meta-analysis considerably increased the number of differentially expressed genes and the number of true discoveries. As previously, the p-value combination method had the best sensitivity, equal to 71.2% with an FDR of 4.6%, higher than the FDR for effect size combination (1.7%), but still below 5%. In terms of gene ranking, the p-value combination also slightly outperformed the other methods with an AUC of 0.966. The limma analysis including a study effect and the moderated effect size approach also performed quite well with AUC of 0.939 and 0.959, respectively. The three methods outperformed a simple limma analysis on the combined expression values, that did not take into account the between-study variability.

Although the IDR criterion has been used by several authors in the literature, it does not check if the additional genes detected with the meta-analyses are actually true positives. In order to compare the different methods we therefore used the number of TP Discoveries. Thus, for the effect size combination method, among the 426 genes detected only with the meta-analysis and not with single study analyses, about 414 were true positives, whereas there were 589 out of 635 with the p-value combination method. This result confirms, as previously observed on sensitivities, that p-value combination outperforms effect size combination. Note that these Discoveries would be even larger if gene-by-gene analyses had been performed for individual studies as in Choi *et al.* (2003), instead of limma analyses.

Table 5. Comparison of global limma analyses - the first one ($Joint_{L1}$) only gathering the expression data, the second one ($Joint_{L2}$) including a study effect in the linear model - with p-value and effect size combinations. The number of differentially expressed genes (DE), FDR, Sensitivity (Sens.), area under ROC curve (AUC), IDR, the number of discoveries (Disc.), IRR and the number of revisions (Revis.) are averaged on 300 simulations, 10 replicates were simulated for both conditions in each study.

	$Joint_{L1}$	$Joint_{L2}$	pv_{Limma}	ES_{Limma}
DE	54.8(9.3)	853.1(19.1)	1064.3(17.7)	732.0(20.2)
Sens.	3.8(0.7)	57.2(1.2)	71.2(1)	50.4(1.3)
FDR	0.0(0.3)	4.3(0.7)	4.6(0.6)	1.7(0.5)
IDR	25.5(6.2)	54.8(1.8)	59.7(1.5)	58.2(1.8)
Disc.	14.1(4.3)	467.2(21.2)	635.1(21.8)	426.4(19.4)
TP Disc.	14.0(4.3)	432.7(18.8)	589.4(19.7)	413.8(18.4)
IRR	91.3(1.5)	17.8(1.6)	8.6(1.2)	34.9(2.1)
Revis.	428.8(18.2)	83.8(9.4)	40.4(6.5)	164(13.2)
TP Revis.	43.3(2.5)	8.2(2.7)	4.0(2.1)	16.3(3.6)
AUC	90.0(0.4)	93.9(0.4)	96.6(0.3)	95.9(0.3)

Concerning the loss of information due to the meta-analysis, among the 470 genes identified by pooling the lists from the individual studies, about 428 genes were dropped on average when jointly analyzing the expression data from the five studies in a simple limma analysis, whereas only about 40 (resp. 164) when combining p-values (resp. effect sizes). The p-value combination method therefore eliminated fewer genes already found by individual studies than the alternative methods. It is, however, interesting to note that the genes which were lost by meta-analysis were mainly false positives. In particular, even if about 40 significant genes found in single analyses were lost by the effect size combination, the true positives among them represented only 4 genes. Therefore, more false positives were lost thanks to the effect size combination than with the p-value combination, which might explain the lower false discovery rate for the former method. The same phenomenon was observed for meta-analysis methods when applying the real application design. Since the number of replicates was large, meta-analysis influenced more accuracy than sensitivity. In this case, IRR was higher than IDR (e.g. with p-value combination: 12.3% vs 9.6%). Concerning the naive global limma analysis, conclusions were completely different from the previous ones. With this setting, there were a lot of discoveries (3457 genes on average) but many of them (3382 genes) were false positives. That explains why the corresponding ROC curve plotted in Figure 3 d. was badly positionned compared to the other ones.

5 DISCUSSION

Extension of shrinkage approaches from moderated t-tests to effect sizes was a natural way to take into account the small sample size in microarray experiments. Thus, not only sensitivity is gained via meta-analysis but also no sensitivity is lost due to the inefficiency of gene-by-gene analyses, especially when there are few replicates. The proposed moderated effect size combination were able to improve traditional effect size meta-analysis approaches. In the comparison study it was found, however, that p-value combination methods usually outperformed effect size combination

approaches. The simulation study showed that in various settings, for different numbers of studies, replicates per study and between-study variability close to the one observed between real prostate cancer datasets, the p-value combination methods outperformed the other meta-analysis methods regarding sensitivity and gene ranking. It is to be noted that for interpretability reasons p-values have to come from the same statistic, and preferably from moderated t-tests such as limma (Smyth, 2004) or SMVar (Jaffrézic *et al.*, 2007). Effect size combination methods were found to be more conservative and offered more accurate results in terms of false positives. A limma analysis including a study effect in the linear model also appeared to be a valuable alternative for meta-analysis. Bayesian meta-analysis methods (Conlon *et al.*, 2007; Scharpf *et al.*, 2007) have not been compared in this simulation study due to very large computing time requirements. However, this does not preclude their potential usefulness in the future.

When comparing the different methods, we found it necessary to report the number of True Discoveries, and not only the number of Discoveries or the Integration Discovery Rate criterion. It is indeed usual to find new genes with meta-analysis, but it has to be checked that they are not false positives. Of course, exact number of True Discoveries can only be known in simulations; for real datasets, only very few genes tend to have their status validated in additional experiments. The number of Revisions or the Integration Revision Rates also have to be considered to evaluate the number of genes lost in the meta-analysis compared to single study analyses. IRR might be higher than IDR in cases where large number of replicates are involved in each individual study.

In this paper, unbiased estimates were given for the proposed moderated effect sizes. Both bias and exact variance of the effect sizes could be calculated with limma and SMVar because both methods give the distribution of the test statistic under the null hypothesis and provide the associated degrees of freedom. The method could not easily be extended to variance modelling papers where the null distribution is not known, such as SAM (Tusher *et al.*, 2001) where the authors use permutations. We also think that the good knowledge of the number of degrees of freedom improved the p-value inverse normal transformations before their combinations and also explains the excellent results of these methods in this paper. In a way, the ability of effect sizes to handle variance components was matched by p-value combination using these moderated t-tests.

With the growing amount of publicly available microarray databases, there will be an increasing interest in combining data from different platforms. Technically our metaMA package allows this integration of different platforms as contrary to the GeneMeta package it can handle missing data. Thus genes not spotted onto some arrays could be treated as missing. Moreover, p-value combination facilitates cross-platform studies. We would however recommend to avoid mixing data from different platforms, if the aim is to increase sensitivity. The minimum we would advise is to only keep genes which could correspond to a common identifier in order to delete missing not at random data. In the case of cross-platform studies, the most difficult job is to match identifiers between platforms; it must be kept in mind that meta-analysis requires a certain data quality (Larsson *et al.*, 2006).

ACKNOWLEDGEMENT

Claus Mayer's work is funded by the Scottish Government Rural and Environment Research and Analysis Directorate (RERAD).

REFERENCES

Choi, J. K., Yu, U., Kim, S., and Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19 Suppl 1**.

Conlon, E. M., Song, J. J., and Liu, A. (2007). Bayesian meta-analysis models for microarray data: a comparative study. *BMC Bioinformatics*, **8**, 80+.

Hedges, L. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* Academic Press.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, **6**(2), 107–128.

Hong, F. and Breitling, R. (2008). A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*.

Hu, P., Greenwood, C. M., and Beyene, J. (2006). Statistical methods for meta-analysis of microarray data: A comparative study. *Inf Syst Front*, **8**, 9–20.

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., and Speed, T. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**(2), 249–264.

Jaffrézic, F., Marot, G., Degrelle, S., Hue, I., and Foulley, J. L. (2007). A structural mixed model for variances in differential gene expression studies. *Genetical research*, **89**(1), 19–25.

Kulinskaya, E. and Staudte, R. G. (2007). Confidence intervals for the standardized effect arising in the comparison of two normal populations. *Statistics in Medicine*, **26**(14), 2853–2871.

Larsson, O., Wennmalm, K., and Sandberg, R. (2006). Comparative microarray analysis. *OMICS*, **10**(3), 381–397.

LaTulippe, E., Satagopan, J., Smith, A., Scher, H., Scardino, P., Reuter, V., and Gerald, W. L. (2002). Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. *Cancer research*, **62**(15), 4499–4506.

Liptak (1958). On the combination of independent tests. *Magyar Tud. Akad. Mat. Kutato Int. Kzl.*, **3**, 171–179.

Loughin, T. M. (2004). A systematic comparison of methods for combining p-values from independent tests. *Computational Statistics & Data Analysis*, **47**(3), 467–485.

Lusa, L., Gentleman, R., and Ruschhaupt, M. (2008). *GeneMeta: MetaAnalysis for High Throughput Experiments*. R package version 1.12.0.

Marot, G. and Mayer, C.-D. (2009). Sequential analysis for microarray data based on sensitivity and meta-analysis. *Stat Appl Genet Mol Biol*, **8**(1).

Rhodes, D. R., Barrette, T. R., Rubin, M. A., Ghosh, D., and Chinnaiyan, A. M. (2002). Meta-analysis of microarrays: Interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res*, **62**(15), 4427–4433.

Scharpf, R. B., Tjelemeland, H., Parmigiani, G., and Nobel, A. B. (2007). A bayesian model for cross-study differential gene expression. *Johns Hopkins University, Dept. of Biostatistics Working Papers*, (Working Paper 158).

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., and Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**(2), 203–209.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, **3**(1).

Stangl, D. K. and Berry, D. A., editors (2000). *Meta-Analysis in Medicine and Health Policy.* Marcel Dekker.

Stevens, J. R. and Doerge, R. W. (2005). Combining affymetrix microarray results. *BMC Bioinformatics*, **6**.

Stouffer, S., Suchman, E., DeVinney, L., Star, S., and Williams, R. J. (1949). The american soldier. adjustment during army life. *Princeton, NJ: Princeton University Press*, **1**.

Stuart, R. O., Wachsman, W., Berry, C. C., Wang-Rodriguez, J., Wasserman, L., Klacansky, I., Masys, D., Arden, K., Goodison, S., McClelland, M., Wang, Y., Sawyers, A., Kalcheva, I., Tarin, D., and Mercola, D. (2004). In silico dissection of cell-type-associated patterns of gene expression in prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(2), 615–620.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, **98**(9), 5116–5121.

4.2 Complementary results

4.2.1 Regarding effect sizes

Effect sizes d are distributed as $(1/\sqrt{\tilde{n}})$ times a non central t-variate with m degrees of freedom ($m = n_i+n_j-2 = n-2$) and noncentrality parameter $\sqrt{\tilde{n}}\delta$ Hedges (1981). For shrinkage effects sizes, m corresponds to the degrees of freedom driven from the associated moderated t-statistics, $\tilde{n} = n_i n_j/(n_i+n_j)$ for limma and $n = n_i + n_j$ for SMVar. From the mean of the noncentral distribution (Johnson and Welch, 1939), we obtain:

$$E(d) = \delta/c(m)$$

with

$$c(m) = \frac{\Gamma(\frac{m}{2})}{\sqrt{\frac{m}{2}}\Gamma(\frac{m-1}{2})}.$$

This leads to equation (13) of the previous paper: the unbiased estimators of effect sizes are: $d' = c(m)d$.

An accurate approximation for c(m) is given in Hedges (1981):

$$c(m) \approx 1 - \frac{3}{4m-1}.$$

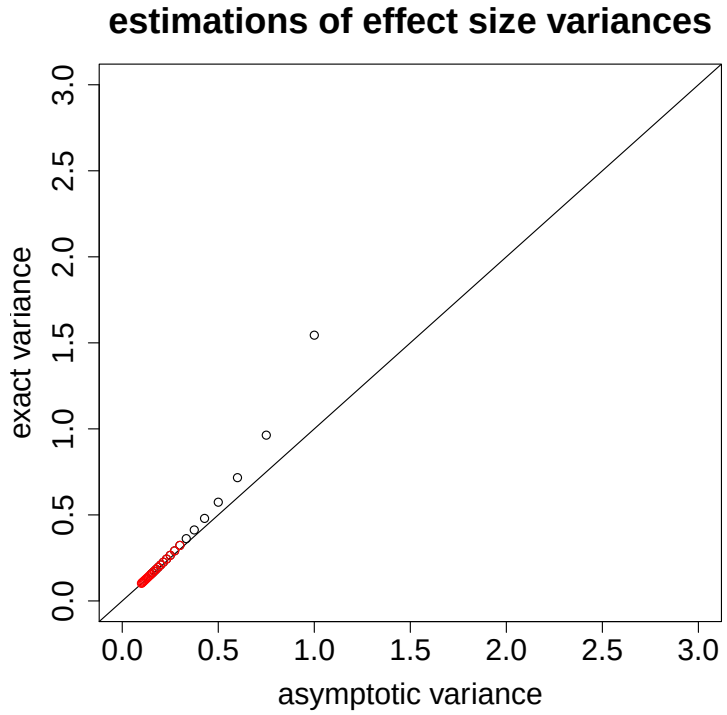
This approximation has a maximum error of 0.007 when $m = 2$, and is accurate to within .00033 when $m \geq 10$.
When $m = n - 2$, we obtain the formula given in Choi *et al.* (2003) and implemented in GeneMeta:

$$d' = d - \frac{3d}{4(n-2)-1}.$$

Note that the variance of the noncentral distribution (Johnson and Welch, 1939) directly provides the exact form of the variance given in equation (10) of our previous paper. Figure 4.1 plots the exact and the asymptotic variances for effect size estimates when the number of replicates per condition n is lower than 30 and $d = 2$.

Red points correspond to $n \geq 10$. To illustrate the difference between results induced by the use of each variance, we artificially split the Singh dataset with 10 replicates in each condition, leaving out the two last prostate tumor samples. We then applied the effect size combination with only changing the variance. Although the difference does not look important for $n = 10$ in Figure 4.1, 116 genes were lost. We can easily imagine the difference people would have with less replicates for which the difference between variances is significant.

Figure 4.1: Exact variance vs. asymptotic variances for effect size estimates



4.2.2 AUC vs. sensitivity

We used both AUC and sensitivity to compare different methods. Indeed, for the same Benjamini-Hochberg threshold, the observed False Discovery Rate in simulations was lower for the effect size combination methods than for the p-value combination methods. Looking at AUC enables to compare sensitivities at similar observed False Discovery Rates. This is of great interest when only gene ranking is interesting. Of course, there is always a price to pay in sensibility when the False Discovery Rate is decreased. Nevertheless, Table 3 in the previous paper emphasizes the limitations of the use of the AUC as a criterion to discriminate between methods. In fact, when inter-study variability was simulated, the AUC was still found to be equal to 89.8% for the naive joint analysis, whereas the sensitivity was extremely poor (equal to 3.9%). This is a very important drawback for practical biological applications. Indeed, even if the real problem comes from the fact that the Benjamini Hochberg correction after effect size combination is too conservative, in practice, when people use R packages, they most of the time choose

the option by default. In metaMA, we choose to use the BH correction. Therefore, sensitivity in simulations might be a better criterion to discriminate between methods if it is compared at the same BH required threshold. It is necessary to check that the corresponding False Discovery Rates are around the required one to avoid to overstate the strength of the results. Since in our case, they were all below or around the required threshold for meta-analyses methods, we clearly advise p-value combination to biologists who use metaMA. Indeed, when they ask for a given FDR, they have more sensitivity while keeping this reasonable FDR with p-value combination than with effect size combination.

In conclusion, when comparing frequentist methods, p-value combination was found to outperform effect size combination methods. This was especially true since we considered moderated tests to calculate p-values. Care has to be taken, however, when calculating p-values. As we saw in the first chapter, variance modelling is really important. In gene expression analysis, the change of the null distribution, for example the use of normal distributions instead of Student distributions and their appropriate number of degrees of freedom has a large impact, leading to losses or gains of hundreds of genes. Concerning the effect size calculations, the between-study variability was here estimated using the method of moments. Effect size combination associated to other methods to estimate between-study variability might give better results than p-value combination. In the next chapter, we will adopt Bayesian approaches to estimate the inter-study variability. In particular, this variability will be modelled in order to enable different inter-study variabilities over conditions.

Chapter 5

Bayesian meta-analysis

Bayesian-based meta-analysis methods for gene expression studies have shown a great promise in the literature (Conlon *et al.* (2007),Scharpf *et al.* (2007)). We did not include any of these methods in the comparison study of the previous chapter due to the very large computing time required for these methods. We were, however, interested in testing at least one of these methods, hoping that Bayesian approach could help estimating more precisely the between-study variance.

5.1 XDE: a Bayesian model for cross-study differential expression

The XDE package is a Bioconductor package and thus belongs to this project gathering many tools for gene expression studies accessible via the R software.

The statistical method implemented in this package is presented in Scharpf *et al.* (2007). While we used to note the expression level y_{gcsr} for gene g, study s, replicate r, and condition c ($c = 1, 2$), we kept the notations of the paper. Thus, y_{gcsr} becomes x_{gsp}, p corresponds to the study while s denotes the sample (replicate). An indicator variable ψ_{sp} indicates to which condition belongs the sample s of study p: $\psi_{sp} \in \{0, 1\}$. A binary parameter δ_g indicates the state of differential expression ($\delta_g = 1$ if the gene is differentially expressed, $\delta_g = 0$ if not). The basic model is written as follows:

$$x_{gsp}|\nu_{gp}, \delta_g, \Delta_{gp}, \sigma^2_{g0p}, \sigma^2_{g1p} \sim \mathcal{N}(\nu_{gp} + \delta_g(2\psi_{sp} - 1)\Delta_{gp}, \sigma^2_{g\psi_{sp}p})$$

Δ_{gp} represents half the average difference between expression levels across phenotypes for gene g in study p.

If the gene is differentially expressed ($\delta_g = 1$), then

$$x_{gsp}|\nu_{gp}, \delta_g = 1, \Delta_{gp}, \sigma^2_{g0p}, \sigma^2_{g1p} \sim \mathcal{N}(\nu_{gp} + \Delta_{gp}, \sigma^2_{g1p})$$

For the next levels of the model specification, we refer to Scharpf *et al.* (2007). Since this Bayesian model involves a lot of parameters, we first used the ones by default. It was impossible for us to run the main function with a dataset of 12625 genes, the R session crashing before the end. Since the use of the package was experimental, we created a subset of 500 genes simulated as in our previous paper with 25 differentially expressed genes. When analysing this subset with metaMA, using the Limma p-value combination method and a 5% Benjamini Hochberg threshold, 23 genes were differentially expressed. All were true positives, only two genes were missing. The IDR equalled to 78% and there was no gene found in individual studies which was lost by meta-analysis. While the results with metaMA were obtained in less than two seconds, running the main function in XDE asking 25000 iterations on the same dataset took a CPU time of 16 hours and 25 minutes. In addition to that, this main function ('xde') stores the MCMC iterations in an object and in written files but does not immediately give the differentially expressed genes over studies. For that, at least two more functions are needed ('calculateBayesianEffectSize' and 'PosteriorAvg', 'Avg' standing for 'Average'). Since R met memory problems when running these last functions, I wrote to Robert Scharpf, the first author of the paper. He answered me very kindly that the problem occured because some of the parameters are indexed by gene, sample, and mcmc iteration and files grow large very quickly. One approach is to set the thinning parameter to 10, which means save every 10th iteration to file. In our specific case, that would mean that posterior averages are only calculated on 2000 iterations since I chose a burn-in of 5001 iterations that is to say that the first 5000 iterations are not taken into account in the calculations. Robert Scharpf also told me that they mostly worked with 3 studies. With 5 studies the covariance matrix for ν_g and Δ_g are much slower mixing. They are working on improvements to the proposals for the covariance matrices that will improve mixing when combining a larger number of studies. As far as I was concerned, I then only kept three studies with 500 genes out of the five I initially simulated. Analysing these 3 studies with metaMA, 19 true positives were found at a 5% Benjamini Hochberg threshold. Once again, there were no false positives and no genes lost due to meta-analysis. IDR was 74%. Using XDE with 25000 iterations comprising a burn-in of 5001 and only saving 1 out of 10 also gave good results. There were 23 genes having a posterior probability to be differentially expressed higher than 0.5. Among them, 22 genes were true positives. The 0.5 threshold was chosen arbitrary, it was kept because its related False Discovery Rate was reasonnable (1/23=4.3%). In other cases or datasets, the threshold could be a higher a posteriori probability. In fact, it is not possible to require a given FDR before the analysis (as it is with metaMA) but several

thresholds have to be tried and then False Discovery Rates calculated. If I similarly process with metaMA and keep the 23 top differentially expressed genes, then 21 genes are true positives. On this example, the Bayesian approach performed slightly better than the frequentist one with the price of a higher computing time. XDE also offers other advantages like the possibility to study discordances (when a gene is up-regulated in one or more studies and down-regulated in others) and concordances between studies and thus detect studies which should be removed from meta-analysis.

5.2 Simplification of the Bayesian hierarchical model

In parallel from the XDE use, we tried to simplify the model in order to gain computing time. The model we proposed for gene g, study s, replicate r, and condition c ($c \in \{1, 2\}$) is the following:

$$y_{gcsr} \sim \mathcal{N}(\mu_{gcs}, \sigma^2_{gcs})$$

$$\mu_{gcs} \sim \mathcal{N}(\theta_g + \delta_g(1 - 2K_c), \tau^2_{gc})$$

K_c is the indicator variable denoting the condition ($K_c = 1$ for condition 1, 0 for condition 2). It has the same role as the previous ψ_{sp}. The following δ must not be confounded with the previous one, which we now prefer to note I since it represents an indicator to know if the gene is differentially expressed. The difference between conditions δ is modelled as follows:

$$\delta_g = I_g \delta^*_g$$

$$I_g \sim Ber(p)$$

(p represents the proportion of differentially expressed genes)

$$\delta^*_g \sim \mathcal{N}(0, \phi^2)$$

$$p \sim beta(a, b)$$

$$\theta_g \sim \mathcal{U}[a_1, a_2]$$

In the next level, all variances follow inverse gamma distibutions, that is to say that the inverse of variances follow a gamma distribution. Parameters were chosen uninformative (0.001 and 0.001). We used either WinBUGS14 or OpenBUGS 3.0.3 to estimate parameters. To avoid 'traps' (Winbugs crashs) after 2000 iterations which were thus not linked to badly specified inits, we

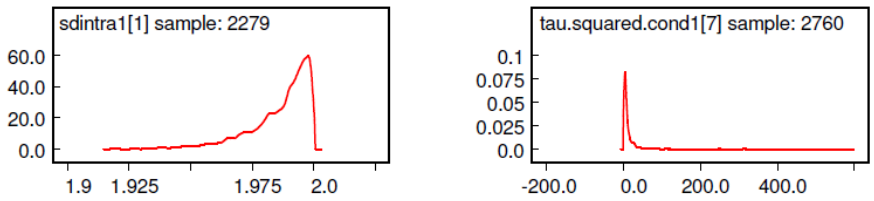
had to shrink within and between variances over genes. For each condition $c \in \{1, 2\}$

$$\ln(\sigma^2_{gcs}) \sim \mathcal{N}(mintra_{cs}, tintra_{cs})$$

$$\ln(\tau^2_{gc}) \sim \mathcal{N}(minter_c, tinter_c)$$

Means *minter* and *mintra* were given normal distributions as priors while standard errors were assumed to follow uniform distributions. We adjusted parameters of the distribution so that they still stay uninformative but without allowing too large variability to avoid traps due to infinite inter-study variances. Figure 5.1 illustrates a bad choice of parameters for between study standard errors.

Figure 5.1: Densities of shrinkage parameters *sdintra* for within-study variability and of inverse of between study variability 'tau.squared.cond2'



When the parameter of the uniform is too small, we clearly see that the a posteriori distribution has been truncated. This leads to infinite variances. Indeed, knowing that in Winbugs, the second parameter of a normal distibution is the precision (inverse of the variance), 'tau.squared.cond2' which is very close to zero represents the inverse of τ^2 which here is infinite for the 7^{th} gene. On the contrary, figure 5.2 reflects good choices of shrinkage parameters: a posteriori distributions differ from a priori distributions. Histories of iterations look correct as shown in the example given at the bottom of the figure.

Despite these good results on shrinkage parameters, we had many difficulties to detect the true differentially expressed genes. Actually, this was not surprising since posterior distributions looked like the prior distributions for p, ϕ^2 and δ^*. Even when forcing p to be very close to the true proportion of simulated differentially expressed genes by decreasing the variance of the beta distribution and leaving its mean equal to the true proportion, δ^* had a posterior distribution similar to the prior one. At least, decreasing the variance of the beta distribution improved the shape of the p autocorrelations.

Figure 5.2: Densities of shrinkage parameters and an example of good history

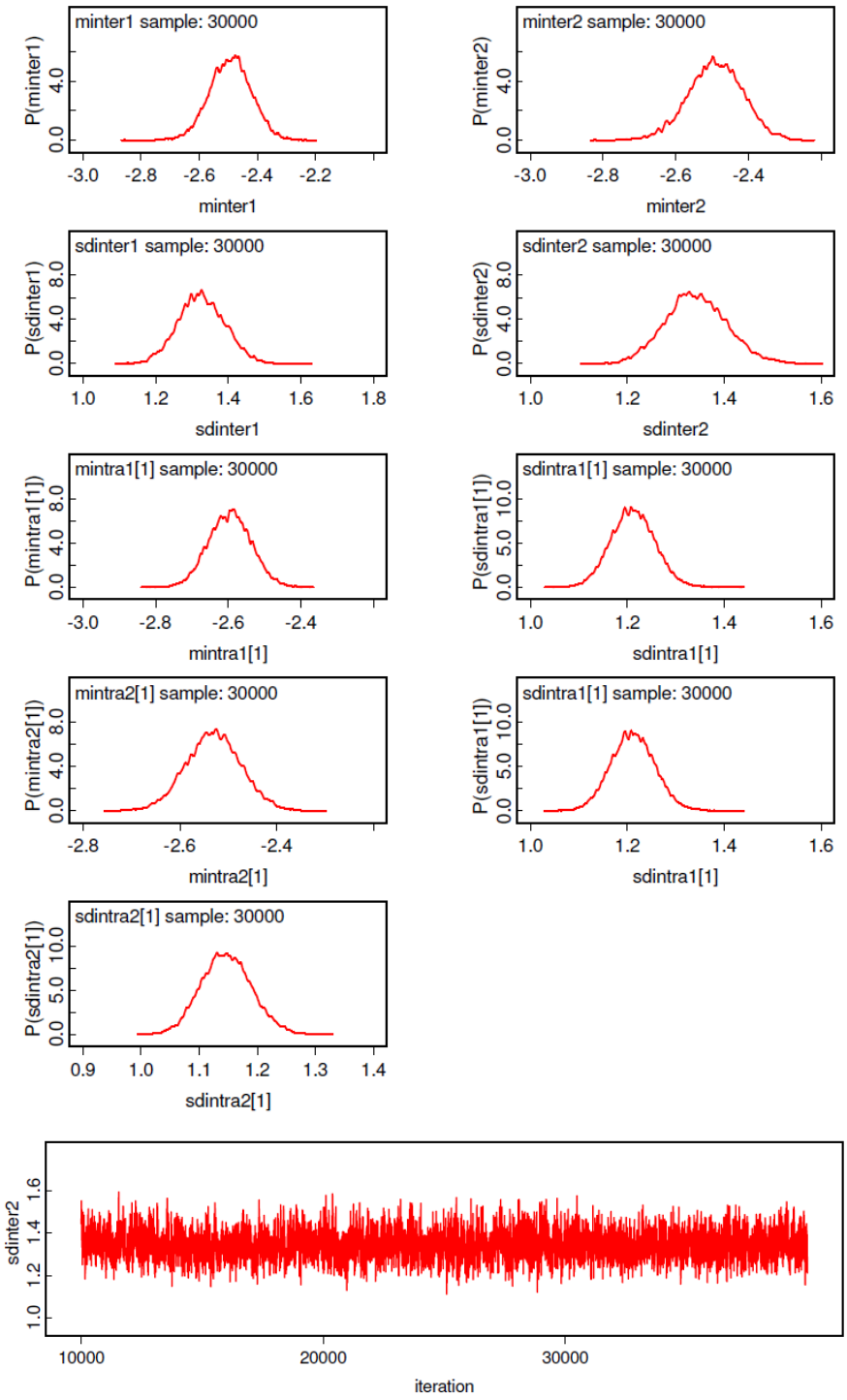
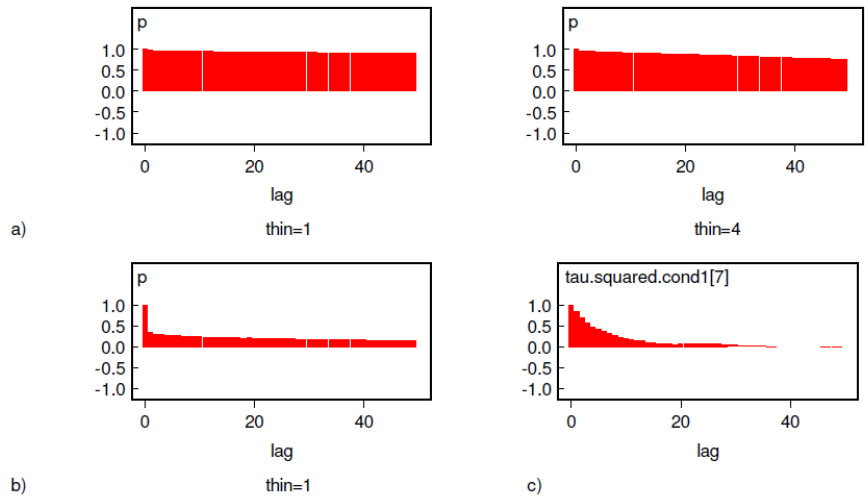


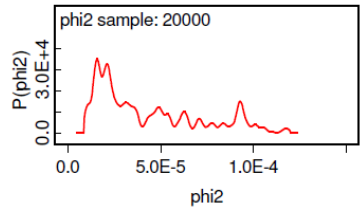
Figure 5.3: Autocorrelations of p and of the inverse of τ^2



On figure 5.3, we can see that a thin of four iterations (only saving one out of four iterations) was not enough (see a)). When decreasing the variance of the beta (see b)), autocorrelations diminish more quickly. The autocorrelations of the inverse of τ^2 (see c)) are given as an example of good shape for autocorrelations. We tried different priors for δ^* (uniform, normal distributions with different variances) and had to eliminate uniform priors leading to WinBUGS traps. We then left normal priors in order to have conjugate laws. Concerning ϕ^2, we first assumed it to be dependent on the gene and each ϕ_g^2 followed an inverse gamma distribution. We also assumed a uniform on the ϕ_g^2 variance, which did not work either. We then assumed one common ϕ^2 for all genes, replacing it by different fixed values. Since the problem was not solved, we tried another model without the indicator variable I and assuming either uniform or log-normal distributions for ϕ^2 which becomes the variance of the δ distribution (δ^* having disappeared at the same time as I). At that point, posterior distributions for ϕ^2 were different from the prior ones (see figure 5.4) but the value found for ϕ^2 was very small, whatever the prior distribution chosen was. This led to bad histories and to strange δ densities.

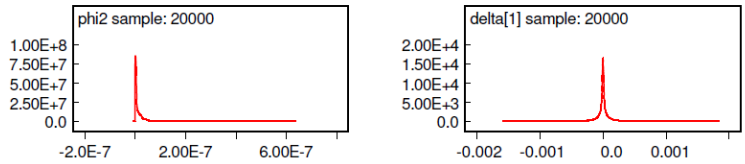
Figure 5.4: Densities, statistic and histories of ϕ^2

log normal distribution on phi2

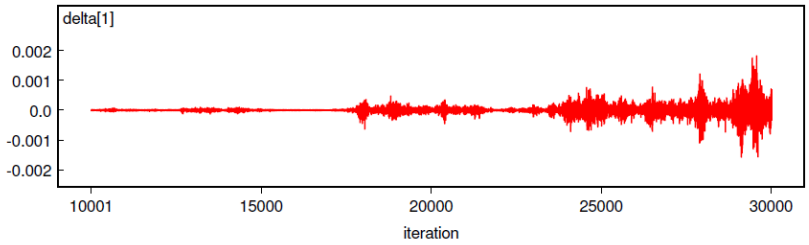


	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
phi2	4.292E-5	2.859E-5	2.397E-6	1.074E-5	3.213E-5	1.04E-4	10001	20000

uniform distribution on phi2

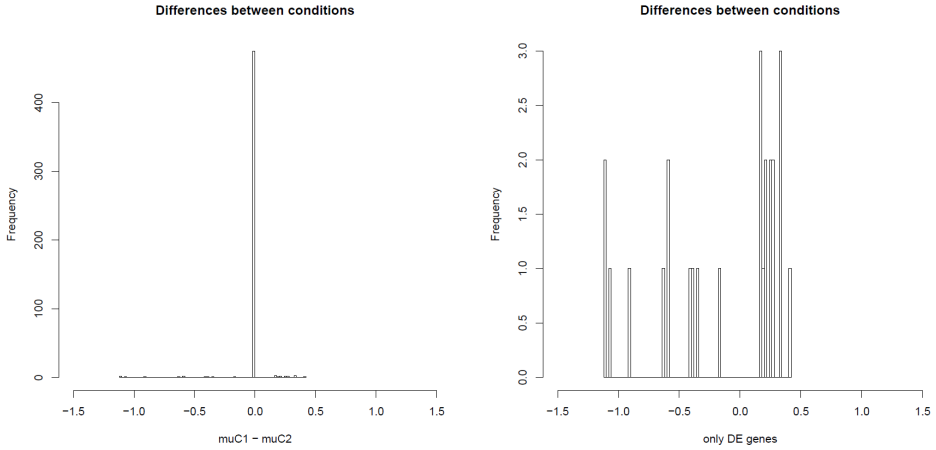


node	mean	sd	MC error	2.5%	median	97.5%	start	sample
phi2	1.884E-8	5.232E-8	3.994E-9	1.383E-11	1.987E-9	1.55E-7	10001	20000



We do not know yet how to improve the model which clearly does not work as it is. We are trying to look in another direction: Bayesian mixture models. Indeed, since 95% of the genes are not differentially expressed, their ϕ^2 value is really 0. The following graph (figure 5.5) represents the true distribution of δ, that is to say the simulated difference between conditions.

Figure 5.5: Simulated differences between conditions



We thus want to include in the hierarchical model 3 distributions in the mixture separating non differentially expressed genes, under and over-expressed genes. This is an outgoing work.

To conclude this chapter, Bayesian analysis proved to be able to perform as well as or even better than frequentist analysis as shown with the use of the XDE package. We thus hope that it will help analysing datasets where a common inter-study variance can not be assumed for both conditions and datasets for which frequentist analysis is not adapted. Bayesian analysis is, however, very time consuming and trying to simplify a complex model also requires expertise and time.

Discussion

Three main approaches have been developed in order to solve the small sample size problem in microarray experiments: variance-covariance modelling, sequential and meta-analysis. The main application considered was the research of differentially expressed genes and we especially concentrated on studies where differences were low and very few differentially expressed genes were expected. Thus, an important criterion used in this PhD was sensitivity, which was more relevant than gene ranking in our applications.

The first part of this PhD proposed shrinkage approaches to model variance-covariance matrices. Shrinkage enables to decrease the number of parameters to estimate - which increases sensitivity - while still keeping a certain flexibility. Information of one particular gene is enriched from information from all other genes. We saw that even if the modelling was completely different from mixture models, it performed similarly to already existing approaches like Varmixt (Delmar *et al.*, 2005). Thanks to the SMVar package which implements the structural model for variances, the approach we proposed has been used not only by biologists we work with but also by people we do not know, as revealed by the questions I received. Concerning the covariance structure modelling, the proposed methods were found to perform well compared to previously proposed empirical Bayesian approaches, and outperformed the gene-specific or common-covariance methods in many cases. The application is, however, not straightforward. In addition to constraints on the number of replicates which must be higher than the number of measures when adopting an empirical bayesian approach, the question about null distibutions and their associated degrees of freedom remains open and turns out to be very important in the context of multiple testing. The paper accepted in CSDA however illustrates the advantage of the use of a structural model on diverse parameters of covariance matrix decompositions when measures are highly correlated.

The second part of this PhD pioneered an innovative approach to microarray experiment designs. While sequential analysis has a long standing tradition in clinical trials in order to reduce costs of the experiments, this

concept had never been introduced for microarrays. Sequential approach enables us to readdress the question of sample size after each stage, which is useful when there is no or little prior knowledge available, which would allow an accurate power calculation. What is more, the stopping rule guarantees to keep a reasonable power when reducing sample size. One interesting feature of the sequential approach for microarrays is that, in contrast to a univariate situation, the large number of genes tested simultaneously prevents the interim analysis from introducing a serious bias to the final p-values. Thus, results from different stages can be combined by meta-analysis methods and error rates can be controlled by applying standard procedures (e.g. the Benjamini-Hochberg rule) to the p-values from the combined stages. We suggested stopping rules based on either the estimated number of true positives or the estimated sensitivity and compared several mixture models that can be used for sensitivity estimation. Our results on simulations and real data sets showed that the application of sequential methods was able to reduce sample-sizes and thus costs in microarray experiments. The hope is then to better organize experiments by including samples saved from some experiments in other ones which would need more samples to increase sensitivity.

The first two parts of this PhD thesis have been linked in the third part on meta-analysis. Indeed, meta-analysis was introduced in sequential analysis in order to combine analyses coming from different stages. In the last part, we extended this meta-analysis to different studies. We especially based our comparison on sensitivity estimates since we assumed that the main problem was the lack of sensitivity in individual studies due to the lack of samples available. In this case, p-value combination outperforms effect size combination. Concerning the gain in accuracy and the loss of false positives, we found that performance in terms of gene ranking was almost similar between moderated effect size and p-value combination. Thus, both methods present similar interest for people who already have enough differentially expressed genes in individual studies and who integrate data in order to provide better predictors rather than increasing sensitivity. In our case, since we sometimes found less than 10 differentially expressed genes per study at a 5 % Benjamini Hochberg rate, we were very interested in sensitivity and thus advise p-value combination. Of course, increasing sensitivity has a cost in terms of false discovery rate which often increases at the same time but our simulation studies checked that FDR stayed below a required threshold. We prefer to combine studies coming from similar platforms to avoid important revision rates caused by difficulties to match identifiers between platforms but do not exclude the possibility to integrate data from different platforms. We also noticed that commercial chips provided better quality than home-made chips

and sometimes the cost paid for the time lost to clean data would be similar to the additionnal cost of commercial chips. We are continuing working on Bayesian analysis in order to provide a better estimation for inter-study variability.

To conclude, although technological progresses have been achieved to produce high quality microarrays, their use for transcriptome analysis still has some limits. While deep sequencing technics are more and more improved and less and less expensive, they offer an alternative to microarrays since they offer many advantages like detection of small RNAs or rare transcripts and non dependence of the genome annotation. Nowadays, they are more used as complementary technics rather than replacement ones but maybe one day, microarrays will become rare experiments. No matter what happens, statistical methods developped in this PhD can be easily fitted to these other types of high dimension biological data.

Bibliography

Albers, C., Jansen, R., Kok, J., Kuipers, O., and van Hijum, S. (2006). Simage: simulation of dna-microarray gene expression data. *BMC Bioinformatics*, **7**, 205.

Anderson, T. (1960). A modification of the sequential probability ratio test to reduce the sample size. *Annals of mathematical statistics*, **31**, 165–197.

Armitage, P. (1960). *Sequential Medical Trials*. Blackwell.

Armitage, P., McPherson, C., and Rowe, B. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society*, **132**, 235–244.

Bauer, P. and Kohne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, **50**, 1029–1041.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, **57**, 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**, 1165–1188.

Brannath, W., Posch, M., and Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association*, **97**, 236–244.

Chang, M. N., Gould, L. A., and Snapinn, S. M. (1995). P-values for group sequential testing. *Biometrika*, **82**(3), 650–654.

Choi, J. K., Yu, U., Kim, S., and Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19 Suppl 1**.

Conlon, E. M., Song, J. J., and Liu, A. (2007). Bayesian meta-analysis models for microarray data: a comparative study. *BMC Bioinformatics*, **8**, 80+.

de Koning, D. J., Jaffrézic, F., Lund, M. S., Watson, M., Channing, C., Hulsegge, I., Pool, M. H., Buitenhuis, B., Hedegaard, J., Hornshøj, H., Jiang, L., Sørensen, P., Marot, G., Delmas, C., Lê Cao, K. A., San Cristobal, M., Baron, M. D.,

Malinverni, R., Stella, A., Brunner, R. M., Seyfert, H. M., Jensen, K., Mouzaki, D., Waddington, D., Jiménez-Marín, A., Pérez-Alegre, M., Pérez-Reinado, E., Closset, R., Detilleux, J. C., Dovc, P., Lavric, M., Nie, H., and Janss, L. (2007). The eadgene microarray data analysis workshop. *Genet Sel Evol*, **39**(6), 621–631.

Degrelle, S. (2006). *Croissance et différenciation du trophoblaste de mammifères en début de gestation: étude par génomique fonctionnelle de l'embryon bovin normal et cloné*. Ph.D. thesis, Université de Versailles-Saint-Quentin-en-Yvelines.

Delmar, P. (2005). *Modèle de mélange sur la variance pour l'analyse différentielle des biopuces*. Ph.D. thesis, Ecole Centrale Paris.

Delmar, P., Robin, S., and Daudin, J. J. (2005). Varmixt: efficient variance modelling for the differential analysis of replicated gene expression data. *Bioinformatics*, **21**(4), 502–508.

Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. M. (1999). Expression profiling using cdna microarrays. *Nature Genetics*, **21**, 10–14.

Guyonnet, B. (2008). *Recherche et identification des gènes différentiellement exprimés dans l'épididyme par une approche transcriptomique. Variations d'expression de ces gènes en relation avec la fertiliité*. Ph.D. thesis, Université Tours.

Hedges, L. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* Academic Press.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, **6**(2), 107–128.

Jaffrézic, F., de Koning, D. J., Boettcher, P. J., Bonnet, A., Buitenhuis, B., Closset, R., Déjean, S., Delmas, C., Detilleux, J. C., Dovc, P., Duval, M., Foulley, J. L., Hedegaard, J., Hornshøj, H., Hulsegge, I., Janss, L., Jensen, K., Jiang, L., Lavric, M., Lê Cao, K. A., Lund, M. S., Malinverni, R., Marot, G., Nie, H., Petzl, W., Pool, M. H., Robert-Granié, C., San Cristobal, M., van Schothorst, E. M., Schuberth, H. J., Sørensen, P., Stella, A., Tosser-Klopp, G., Waddington, D., Watson, M., Yang, W., Zerbe, H., and Seyfert, H. M. (2007). Analysis of the real eadgene data set: Comparison of methods and guidelines for data normalisation and selection of differentially expressed genes. *Genet Sel Evol*, **39**(6), 633–650.

Jaffrézic, F., Marot, G., Degrelle, S., Hue, I., and Foulley, J.-L. (2007). A structural mixed model for variances in differential gene expression studies. *Genetical Research*, **89**(1), 19–25.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, **1**, 361–379.

Johnson, N. and Welch, B. (1939). Applications of the noncentral-t-distribution. *Biometrika*, **31**, 362–389.

Kerr, M., Afshari, C., Bennett, L., Bushel, P., Martinez, J., Walker, N., and Churchill, G. (2002). Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica*, **12**, 203–217.

Kieser, M., Bauer, P., and Lehmacher, W. (1999). Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biometrical Journal*, **41**(3), 261–277.

Lan, K. and DeMets, D. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, **70**, 659–663.

Lê Cao, K.-A. (2008). *Transcriptome : outils statistiques pour l'étude cinétique et l'intégration de variables biologiques ou métabolomiques*. Ph.D. thesis, Université Toulouse.

Lee, J. W. (1994). Group sequential testing in clinical trials with multivariate observations: a review. *Statistics in Medicine*, **13**(2), 101–111.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, **55**(4), 1286–1290.

Marot, G. and Mayer, C.-D. (2009). Sequential analysis for microarray data based on sensitivity and meta-analysis. *Stat Appl Genet Mol Biol*, **8**(1).

Mary-Huard, T. (2006). *Réduction de la dimension et sélection de modèles en classification supervisée*. Ph.D. thesis, Université Paris XI.

Mary-Huard, T., Picard, F., and Robin, S. (2006). *Mathematical and Computational Methods in Biology. Chapter Introduction to Statistical Methods for Microarray Data Analysis*. Hermann.

McLachlan, G. J., Bean, R. W., and Jones, L. B. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, **22**(13), 1608–1615.

Neuvial, P. (2008). *Contributions à l'analyse statistique des données de puces à ADN*. Ph.D. thesis, Université Paris VII.

O'Brien, P.-C. and Fleming, T.-R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, **35**, 549–556.

Peto, R., Pike, M., and Armitage, P. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient . i. introduction and design. *British Journal of Cancer*, **34**, 585–612.

Pocock, S.-J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, **64**, 191–199.

Robin, S., Barhen, A., Daudin, J., and Pierre, L. (2007). A semi-parametric approach for mixture models: Application to local false discovery rate estimation. *Computational Statistics and Data Analysis*, **51**(12), 5483–5493.

Schäfer, H., Timmesfeld, N., and Müller, H.-H. (2006). An overview of statistical approaches for adaptive designs and design modifications. *Biometrical Journal*, **48**(4), 507–520.

Scharpf, R. B., Tjelemeland, H., Parmigiani, G., and Nobel, A. B. (2007). A bayesian model for cross-study differential gene expression. *Johns Hopkins University, Dept. of Biostatistics Working Papers*, (Working Paper 158).

Sébille, V. and Bellissant, E. (2003). Sequential methods and group sequential designs for comparative clinical trials. *Fundamental and Clinical Pharmacology*, **17**(5), 505–516.

Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**(1).

Speed, T. (2003). *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall CRC.

Spiessens, B., Lesaffre, E., Verbeke, G., Kim, K., and DeMets, D. L. (2000). An overview of group sequential methods in longitudinal clinical trials. *Statistical Methods in Medical Research*, **9**(5), 497–515.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences of the United States*, **100**, 9440–9445.

Tang, D.-I., Gnecco, C., and Geller, N. L. (1989). Design of group sequential clinical trials with multiple endpoints. *Journal of the American Statistical Association*, **84**(407), 776–779.

Todd, S. (2007). A 25-year review of sequential methodology in clinical studies. *Statistics in Medicine*, **26**(2), 237–252.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, **98**(9), 5116–5121.

Victor, A. and Hommel, G. (2007). Combining adaptive designs with control of the false discovery rate–a generalized definition for a global p-value. *Biometrical Journal*, **49**(1), 94–106.

Wald, A. (1947). *Sequential analysis*. Wiley.

Watson, M., Pérez-Alegre, M., Baron, M. D., Delmas, C., Dovc, P., Duval, M., Foulley, J. L., Garrido-Pavón, J. J., Hulsegge, I., Jaffrézic, F., Jiménez-Marín, A., Lavric, M., Lê Cao, K. A., Marot, G., Mouzaki, D., Pool, M. H., Robert-Granié, C., San Cristobal, M., Tosser-Klopp, G., Waddington, D., and de Koning, D. J. (2007). Analysis of a simulated microarray dataset: Comparison of methods for data normalisation and detection of differential expression. *Genet Sel Evol*, **39**(6), 669–683.

Wei, L. J., Su, J. Q., and Lachin, J. M. (1990). Interim analyses with repeated measurements in a sequential clinical trial. *Biometrika*, **77**(2), 359–364.

Whitehead, J. and Stratton, I. (1983). Group sequential clinical trials with triangular continuation regions. *Biometrics*, **39**, 227–236.

Wit, E. and McClure, J. (2004). *Statistics for Microarrays: Design, Analysis and Inference*. Wiley.

Xiong, C., Yu, K., Gao, F., Yan, Y., and Zhang, Z. (2005). Power and sample size for clinical trials when efficacy is required in multiple endpoints: application to an alzheimer's treatment trial. *Clin Trials*, **2**(5), 387–393.

Yang, Y., Dudoit, S., Luu, P., Peng, V., Ngai, J., and Speed, T. (2002). Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, **30**.

Appendices

Glossary

Alternative splicing: Process by which the exons of the RNA produced by transcription of a gene are reconnected in multiple ways during RNA splicing. The resulting different mRNAs may lead to different proteins.

CGH (Comparative genomic hybridization): Method for the analysis of copy number changes (gains/losses) in the DNA content of a given subject's DNA and often in tumor cells. Array CGH detects genomic copy number variations at a higher resolution level than chromosome-based CGH.

ChIP on chip: Chromatin immunoprecipitation on Chip. Technique used to investigate interactions between proteins and DNA in vivo.

Copy number variation (CNV): Segment of DNA in which copy-number differences have been found by comparison of two or more genomes.

Exon: DNA region within a gene that is translated into protein.

Fluorophore: Component which causes a molecule to be fluorescent.

Functional genomic: Study of gene functions, and their expression regulation and interactions.

Hybridization: Recognition and interaction of two complementary sequences.

Intron: DNA region within a gene that is not translated into protein.

Metabolome: Collection of all the organic compounds in a biological organism.

mRNA: messenger ribonucleic acid. Macromolecule formed by a single helical strand of similar structure as one of the two strands which constitute DNA. RNA differs from DNA with the replacement of a sugar, the deoxyribose, by another one, the ribose and the replacement of a nucleobase, thymine, by uracil.

Polymerase Chain Reaction: Amplification technique to increase DNA quantity from a given sample.

Proteome: Collection of all the proteins of an organism.

RNA splicing: Modification of an RNA after transcription, in which introns are removed and exons are joined.

SNP: Single Nucleotide polymorphism. DNA sequence variation occurring when a single nucleotide in the genome (or other shared sequence) differs between members of a species (or between paired chromosomes in an individual).

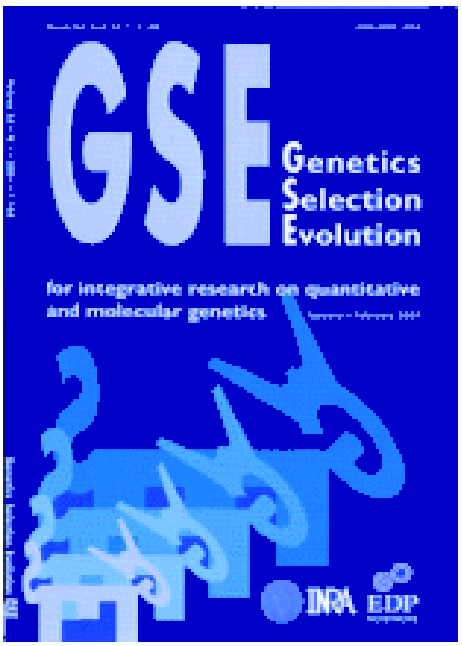
Tiling arrays: Genome tiling arrays consist of overlapping probes designed to densely represent a genomic region of interest, sometimes as large as an entire human chromosome. The purpose is to empirically detect expression of transcripts or alternatively splice forms which may not have been previously known or predicted.

Traduction: Synthesis of a protein from a mRNA.

Transcription: Process by which the nucleotides sequence of a gene is copied as one strand of RNA.

Transcriptome: Collection of all mRNA.

Papers written after EADGENE workshop



Genet. Sel. Evol. 39 (2007) 621–631
© INRA, EDP Sciences, 2007
DOI: 10.1051/gse:2007028

Available online at:
www.gse-journal.org

Original article

The EADGENE Microarray Data Analysis Workshop (*Open Access publication*)

Dirk-Jan DE KONING[a*], Florence JAFFRÉZIC[b], Mogens Sandø LUND[c], Michael WATSON[d], Caroline CHANNING[a], Ina HULSEGGE[e], Marco H. POOL[e], Bart BUITENHUIS[c], Jakob HEDEGAARD[c], Henrik HORNSHØJ[c], Li JIANG[c], Peter SØRENSEN[c], Guillemette MAROT[b], Céline DELMAS[f], Kim-Anh LÊ CAO[f,g], Magali SAN CRISTOBAL[f], Michael D. BARON[h], Roberto MALINVERNI[i], Alessandra STELLA[i], Ronald M. BRUNNER[j], Hans-Martin SEYFERT[j], Kirsty JENSEN[a], Daphne MOUZAKI[a], David WADDINGTON[a], Ángeles JIMÉNEZ-MARÍN[k], Mónica PÉREZ-ALEGRE[k], Eva PÉREZ-REINADO[k], Rodrigue CLOSSET[l], Johanne C. DETILLEUX[l], Peter DOVČ[m], Miha LAVRIČ[m], Haisheng NIE[n], Luc JANSS[c]

[a] Roslin Institute, Roslin, UK
[b] INRA, UR337, Jouy-en-Josas, France
[c] University of Aarhus, Tjele, Denmark
[d] Institute for Animal Health, Compton, UK
[e] Animal Sciences Group Wageningen UR, Lelystad, The Netherlands
[f] INRA, UMR444, Castanet-Tolosan, France
[g] Université Paul Sabatier, Toulouse, France
[h] Institute for Animal Health, Pirbright, UK
[i] Parco Tecnologico Padano (PTP), Lodi, Italy
[j] Research Institute for the Biology of Farm Animals, Dummerstorf, Germany
[k] University of Córdoba, Córdoba, Spain
[l] University of Liege, Liege, Belgium
[m] University of Ljubljana, Ljubljana, Slovenia
[n] Wageningen University and Research Centre, Wageningen, The Netherlands

(Received 10 May 2007; accepted 3 July 2007)

Abstract – Microarray analyses have become an important tool in animal genomics. While their use is becoming widespread, there is still a lot of ongoing research regarding the analysis of microarray data. In the context of a European Network of Excellence, 31 researchers representing 14 research groups from 10 countries performed and discussed the statistical analyses

* Corresponding author: DJ.dekoning@bbsrc.ac.uk

of real and simulated 2-colour microarray data that were distributed among participants. The real data consisted of 48 microarrays from a disease challenge experiment in dairy cattle, while the simulated data consisted of 10 microarrays from a direct comparison of two treatments (dye-balanced). While there was broader agreement with regards to methods of microarray normalisation and significance testing, there were major differences with regards to quality control. The quality control approaches varied from none, through using statistical weights, to omitting a large number of spots or omitting entire slides. Surprisingly, these very different approaches gave quite similar results when applied to the simulated data, although not all participating groups analysed both real and simulated data. The workshop was very successful in facilitating interaction between scientists with a diverse background but a common interest in microarray analyses.

gene expression / two colour microarray / statistical analysis

1. INTRODUCTION

The recent development of high throughput gene-expression technologies, such as microarrays, has given rise to a plethora of new research hypotheses and possibilities. Extensive reviews are available about the application [3], design [6], and analysis [12] of microarray studies. In livestock, microarrays have been proposed to study gene-expression in the parasite (Malaria [13]; Trypanosomosis [8]) as well as host response following infection (*e.g. Mycobacterium paratuberculosis* infection in cattle [7]; *Eimeria* infection in poultry [11]). Other applications in livestock include the evaluation of the effects of diet on gene expression in beef cattle [4] and gene expression differences related to differences of muscling in pigs [5].

In a recent review, Allison *et al.* outline the areas of consensus and outstanding questions with regards to microarray analysis [2]. Some points of consensus regarding data analysis as presented by those authors [2] were the following: (1) many methods exist for the pre-processing (normalisation, *etc.*) of two-colour microarrays, but there is no clear winner and none were discussed in detail; (2) using fold-change alone as a test for differential expression is inefficient; (3) false discovery rate is a good alternative to conventional multiple testing; and (4) unsupervised classification is overused and should be validated using re-sampling techniques. The most relevant outstanding questions were [2] the following: (1) the best image processing algorithm; (2) the evaluation of data quality; and (3) the assessment of intersections between sets of findings within and between experiments.

Given the lack of consensus in many areas, especially for the two-colour arrays that are abundant in livestock research, we organised a workshop on the analysis of microarrays. Conferences dealing with the statistical analyses of

microarrays using common sets of data have been successfully organised annually in the United States since 2000 [10] (http://www.camda.duke.edu/). These conferences have been large scale events, attracting 250 or more participants. In contrast, the present workshop was limited to 35 participants to maximise interaction and focussed on microarray experiments in the context of the genetics of host-pathogen interaction in livestock. The workshop was organised through the EC-funded network of excellence (NoE) EADGENE (European Animal Disease Genetics Network of Excellence for Animal Health and Food Safety; http://www.eadgene.info/).

2. WORKSHOP GOALS

The main aim of the workshop was to bring together scientists from within the EADGENE network with an interest in microarray analyses and to facilitate interaction and future collaboration between these scientists. In order to focus the discussions, the workshop was organised around two sets of data, real and simulated, that were distributed among the participants prior to the workshop. The methods of analysis, the interpretation of results and how to use the (quite complex) real experimental design were left to the participants. This was advantageous as it led to very different approaches by different groups. The diversity of approaches was a major contributor to our ability to identify outstanding questions in the treatment of microarray data.

The statistical aspects of a microarray study include the design of the study, the quantification of the hybridisation intensities, the pre-processing and normalisation of data, the inference and classification of results, the biological interpretation and finally the validation of differentially expressed genes as well as other follow-up studies. For practical reasons this workshop only dealt with the following aspects of microarray analysis: (1) some of the pre-processing of the raw microarray intensities (mainly quality control); (2) normalisation of the microarray data; (3) the detection of differentially expressed genes; (4) the clustering and classification analyses of the differentially expressed genes as well as the biological interpretation (real data only).

The workshop format allowed comparison of results for a real microarray experiment that was relevant to the remit of EADGENE as well as simulated data with known parameters, which facilitated a comparison of performance between groups. However, it must be stressed that the interaction among scientists, facilitated through common data sets, was the main objective.

Table I. Overview of workshop participants and acronyms used to describe the groups throughout the four articles.

Acronym	Affiliation	Group size	Real data	Simulated data
AARHUS	University of Aarhus, Tjele, Denmark	6	x	
CDB	University of Córdoba, Spain	3		x
IAH_C	Institute for Animal Health, Compton, UK	1	x	
IAH_P	Institute for Animal Health, Pirbright, UK	1		x
IDL	Animal Sciences Group Wageningen UR, Lelystad, NL	2	x	x
INRA_J	INRA, Jouy-en-Josas, France	3	x	x
INRA_T	INRA, Castanet-Tolosan, France	(8, 6)[a]	x (8)	x (6)
ULg	University of Liege, Liege, Belgium	2 x 1[b]	x	
PTP	Parco Tecnologico Padano, Lodi, Italy	3	x	
RIBFA	Research Institute for the Biology of Farm Animals, Dummerstorf, Germany	5[c]	x	
ROSLIN	Roslin Institute, Roslin, UK	4	x	x
SLN	University of Ljubljana, Slovenia	2	x	x
WUR	Wageningen University and Research Centre, NL	2	x	

[a] This group included six members for the analysis of simulated data (four from INRA-Station d'amélioration génétique des animaux and two from INRA-Laboratoire de génétique cellulaire) and eight members for the analysis of real data (four from INRA-Station d'amélioration génétique des animaux, three from INRA-Laboratoire de génétique cellulaire and one from the Paul Sabatier University).

[b] Two members from different groups within the University of Liege analysed and presented independently.

[c] This group included two members from Ludwig-Maximilian University in München, and one member of the University of Veterinary Medicine in Hannover.

3. THE WORKSHOP PARTICIPANTS

The data was analysed by 42 participants, representing 14 research groups from 11 EADGENE partners. During a 3-day workshop, attended by 31 participants, all groups presented and discussed their findings. The details of the different groups as well as their acronym and group sizes are presented in Table I. While all participants had shared interests through their involvement in EADGENE, they had varying levels of experience in the analyses of microarray data and different interests in taking part.

Some participants were routinely involved in the analyses of microarrays in their own institutes while others were using this workshop to gain 'hands-on' experience with the analyses of microarray data. Some groups had developed sophisticated tools to deal with a specific aspect of microarray analyses and used the workshop to demonstrate or test-drive their approach. Because the

real data was from a mastitis experiment, some participants had a particular interest in this disease and its study *via* microarray analyses.

The detailed results on the analyses of the real data are given by Jaffrézic *et al.* [9] for the quality control, the normalisation and statistical testing and Sørensen *et al.* [14] for the multiple gene analyses. The detailed results of the simulated data analyses are presented by Watson *et al.* [15].

4. OVERVIEW OF DATA

4.1. Real data

The real data consisted of 48 microarrays from an artificial infection experiment in dairy cattle with several time points and two different infectious agents: *Escherichia coli* and *Staphylococcus aureus*.

For further details on the experimental procedures see Jaffrezic *et al.* [9].

The microarray experiment was carried out using the Bovine 20K array (ARK-Genomics: http://www.ark-genomics.org/). A reference design, without dye-swap, was used and the reference sample was made up of a pool of all 48 RNA samples. The resulting microarrays were scanned and data were extracted using BlueFuse (BlueGnome, http://www.cambridgebluegnome.com/bluefuse.htm). No further adjustments or normalisations were made to these data prior to distribution to the participants. The distributed data included an automated annotation of the microarray provided by Mark Fell (Roslin Institute).

4.2. Simulated data

The microarray data were simulated using Simage [1] (http://bioinformatics.biol.rug.nl/websoftware/simage/simage_start.php). This provides a menu-driven interface in which the user can define gene effects as well as numerous noise factors. Using "Simage-R Parameter" we estimated summary statistics from a randomly selected microarray slide from the real data and used this to simulate 10 slides of a direct comparison (A *versus* B) where every second slide had treatments reversed for dyes. Although the parameter settings for the simulated data were derived from a real microarray, a lot of noise was added to really test the QC and normalisation approaches of the various groups. From the 2400 genes that were simulated, 624 were differentially expressed (264 up regulated from A to B and 360 down regulated).

4.3. Differences between real and simulated data

While the simulated data provided a simple A *versus* B comparison, the real data had the components of time and type of bacteria that were used to infect the cows. The researcher could ask different questions: *e.g.* which genes are differentially expressed following infection with a specific pathogen? At what time post infection is the differential expression most prominent? What genes are only differentially expressed following infection with one pathogen and not the other? Although unintended as a discussion area for the workshop, the real data did illustrate a problem in experimental design, which was not discovered by a pilot experiment, namely the dependence in gene expression between the udder quarters in the real data.

The results from the workshop may at first glance seem contradictory: the real data results were quite different between groups (both in numbers of differentially expressed genes and gene order) [9] while the results from the simulated data indicate that most of the approaches gave good and comparable results [15]. It could be argued that methods that give similar results for simulated data should give similar results for real data, if the two sets of data are comparable. The difference between the results from real and simulated data is most likely due to differences in the expected statistical power to detect differentially expressed genes between the two sets of data: while the real data consisted of 48 microarrays, any comparison between two time points within an infection had only four microarrays contributing to each time point, while the contrast was indirect *via* a reference design. The simulated data consisted of 10 microarrays for a direct A *versus* B comparison making it more powerful than the comparisons within the real data. It could be argued retrospectively that for comparison of methods the real data had too little power and too many possible scenarios to be tested, while the simulated data had too much power to reveal subtle differences between methods. This emphasises the benefits of this kind of workshop, since this finding was only apparent after combining and contrasting the approaches and results of the different groups, and the observation will be fed forward into future workshops.

In the simulated data only two levels of differential expression were simulated: one for up-regulated genes and one for the down-regulated genes. Even so, the mixture model distributions of test statistics showed that the various noise contributions produced symmetrical distributions for the up and down regulated genes. The simulated data was notably different from real data [15], but still allowed valuable comparisons of approaches to analysis. The simulated data did confirm that when the power of an experiment is high, many of the specific differences between the methods that are applied may become

less crucial, provided that they deal adequately with high levels of technical bias or noise. At the same time, many microarray experiments have moderate to low power and hence comparison of methods on the basis of real data has considerable merit.

5. QUESTIONS, CONSENSUS, AND RECOMMENDATIONS

5.1. Quality Control (QC)

The approaches presented during the workshop showed most divergence at the QC stage. In terms of QC of the real data, several groups used the spot quality indicators provided by the scanning software (Bluefuse) to make decisions about excluding spots from the analysis. Other groups indicated that they would normally take account of background intensities for quality control but Bluefuse does not use a measure of background intensity from pixels around the spot, nor does it make an explicit estimate from within-spot pixels, and therefore the background intensities were not provided for the real data. One group re-estimated background from the data provided and used ratios between signal and inferred background to exclude bad spots while another group excluded spots on the basis of absolute intensity (mainly for the simulated data). Further to omitting bad spots based on quality indicators or (relative) intensity, some groups omitted entire slides from further analyses based on QC criteria. As an alternative to using quality indicators to include or omit spots from further analyses, one group used quality indicators as statistical weights in both normalisation and analysis of the microarrays.

The different approaches for QC led to some groups omitting no spots at all while other groups omitted many spots and even entire slides (up to two or three slides for the simulated data). Removing spots from subsequent analyses often renders the statistical model unbalanced and reduces the degrees of freedom, and hence the power of the test for a single gene. The effect of removing spots on normalisation, shrinkage of gene variance and multiple testing may counterbalance the loss of power, but these effects are less predictable. Therefore, approaches that utilise all spots but account for different quality of spots deserve further attention.

Another point of discussion was when to apply the QC: It was argued that outlier spots can only be identified after normalisation has taken account of spatial effects but the counter argument was that spots with saturated intensity measurements would bias the normalisation and should be removed before normalisation. It was suggested that QC should be applied at several stages of

the analyses but this was not implemented by any of the groups. Although QC was widely debated during the workshop, we did not define a 'best practise' for QC, although we can make a recommendation to evaluate the effect of various levels of spot editing. Again, the benefits of the workshop are shown by the unexpected identification of QC and data editing as critical factors for discussion and further study. Many publications concentrate on the statistical analysis of simulated or established datasets. While comparison showed that the statistics are generally well understood or accepted; how real experimental data is pre-processed remains a matter for further study.

5.2. Normalisation, significance testing and multi-gene analyses

For the normalisations, many groups removed intensity related bias (when the relationship between average intensity and the ratio between the two colour intensities is non-linear) by LOWESS (or LOESS) regression. One group did additional spatial smoothing while few groups included across-slide normalisation. The latter is emphasised in the results for the simulated data [15] as a way of making slides more comparable, especially for the noisy data in this study. The gene expression contrasts were mainly estimated using linear models or mixed linear models with various approaches to shrink the gene variance prior to significance testing. To address the multiple testing issues, most groups used some variant of the false discovery rate (FDR) but there were also some standard and novel approaches based on mixture distributions. The main problem of comparing gene lists between groups is that it could not be determined what stage of the different analysis pipelines caused the results to differ.

The only three approaches that performed very poorly in detecting differentially expressed genes in the simulated data was one approach based on fold-changes only, and two using ANOVA combined with the lowest level of normalisation – chip median correction. Because of the prominent print-tip effects in the simulated microarray data, failure to account for these will result in many spurious effects when using only chip-median correction and/or analysing fold-changes only [15].

With regards to the recent review by Allison *et al.* [2], the workshop echoed the points regarding the current lack of agreement in pre-processing (although the main differences were in QC rather than normalisation), in particular that fold-changes are not good criteria, and that the FDR, as well as some other novel multiple testing approaches, provide an attractive alternative to conventional multiple testing strategies. We did not address the outstanding questions

on image processing algorithms because we only used results from a single processing algorithm for the workshop.

The multi-gene analyses were too diverse for a meaningful comparison although some trends are described by Sørensen *et al.* [14]. Those analyses that were aimed at assigning biological meaning to the differentially expressed genes were hampered by the limited annotation that was available for the clones on the microarray. This will improve over time with the ongoing annotation of the cow genome sequence.

5.3. Recommendations

While the participants agreed that the workshop had been very useful, we also debated recommendations for potential future workshops on the same topic. The following recommendations were made: (1) Provide different levels of pre-processed data for different analyses. You can provide raw image files to compare image processing algorithms, while you also provide normalised data to compare different models to obtain gene lists or to compare different clustering approaches. Likewise, when comparing bioinformatics tools for the biological interpretation of microarray results, you provide a pre-set common gene list, preferably for a model species with good bioinformatics resources. (2) Ask participants to analyse specific contrasts or scenarios. For the present workshop, we gave real data on 48 slides as well as experimental details, but the participants could decide on what part of the data they would use and what contrasts they would estimate. (3) Simulate microarrays that are more similar to real data and if possible, include a range of gene effects and variances as well as a correlation structure among genes. (4) Because of limited presentation time at the workshop, some details of the analyses were missed, in particular analyses that were initially done, but not carried further. One option is to have a pre-meeting participant survey that includes what approaches were tested and how they performed. Other ways of having a more uniform reporting structure among groups may also benefit the comparisons.

6. CONCLUSIONS

The workshop succeeded in its main aim of sharing expertise and experience among statisticians and biologists using microarrays in livestock research. At least one group used it as a starting point to re-visit their own data analysis pipeline in the view of analyses and expectations of other groups. Furthermore, the three companion papers with details on the various analyses and results will

provide pointers for colleagues in the wider community regarding the options available for microarray analyses [9, 14, 15].

While a direct comparison of results between groups remained challenging, it was extremely useful to discuss microarray analyses on the basis of two common data sets. In many conferences or workshops, participants only present their own data and hence any conclusions about methods cannot be separated from the experiments to which these methods were applied. The joint analyses of the same data that was done during the workshop will also have added value to the original experiment. Furthermore, the workshop is not an endpoint but the starting point of new collaborations among researchers analysing microarray data but also between these researchers and biologists that ultimately give meaning to the results.

ACKNOWLEDGEMENTS

The organisers acknowledge local support from the University of Aarhus colleagues, in particular Karin Smedegard. We greatly appreciate the data contributed to the workshop by Hans-Martin Seyfert, the analysis done by Kirsty Jensen and all the work included of their colleagues. For the simulated data, we acknowledge early access to Simage thanks to Ritsert Jansen and his group. The workshop was organised and financially supported by EADGENE (EU Contract No. FOOD-CT-2004-506416).

REFERENCES

[1] Albers C.J., Jansen R.C., Kok J., Kuipers O.P., van Hijum S.A., SIMAGE: simulation of DNA-microarray gene expression data, BMC Bioinformatics 7 (2006) 205.

[2] Allison D.B., Cui X., Page G.P., Sabripour M., Microarray data analysis: from disarray to consolidation and consensus, Nat. Rev. Genet. 7 (2006) 55–65.

[3] Butte A., The use and analysis of microarray data, Nat. Rev. Drug Discov. 1 (2002) 951–960.

[4] Byrne K.A., Wang Y.H., Lehnert S.A., Harper G.S., McWilliam S.M., Bruce H.L., Reverter A., Gene expression profiling of muscle tissue in Brahman steers during nutritional restriction, J. Anim. Sci. 83 (2005) 1–12.

[5] Cagnazzo M., te Pas M.F., Priem J., de Wit A.A., Pool M.H., Davoli R., Russo V., Comparison of prenatal muscle tissue expression profiles of two pig breeds differing in muscle characteristics, J. Anim. Sci. 84 (2006) 1–10.

[6] Churchill G.A., Fundamentals of experimental design for cDNA microarrays, Nat. Genet. 32 (Suppl.) (2002) 490–495.

[7] Coussens P.M., Jeffers A., Colvin C., Rapid and transient activation of gene expression in peripheral blood mononuclear cells from Johne's disease positive cows exposed to *Mycobacterium paratuberculosis* in vitro, Microb. Pathog. 36 (2004) 93–108.

[8] El Sayed N.M., Hegde P., Quackenbush J., Melville S.E., Donelson J.E., The African trypanosome genome, Int. J. Parasitol. 30 (2000) 329–345.

[9] Jaffrézic F., de Koning D.J., Boettcher P.J., Bonnet A., Buitenhuis B., Closset R., Déjean S., Delmas C., Detilleux J.C., Dovč P., Duval M., Foulley J.-L., Hedegaard J., Hornshoj H., Hulsegge I., Janss L., Jensen K., Jiang L., Lavrič M., Lê Cao K.-A., Lund M.S., Malinverni R., Marot G., Nie H., Petzl W., Pool M.H., Robert-Granié C., San Cristobal M., van Schotshorst E.M., Schuberth H.-J., Sørensen P., Stella A., Tosser-Klopp G., Waddington D., Watson M., Yang W., Zerbe H., Seyfert H.-M., Analysis of the real EADGENE data set: Comparison of methods and guidelines for data normalisation and selection of differentially expressed genes., Genet. Sel. Evol. 39 (2007) 633–650.

[10] Johnson K., Lin S., Call to work together on microarray data analysis, Nature 411 (2001) 885.

[11] Min W., Lillehoj H.S., Kim S., Zhu J.J., Beard H., Alkharouf N., Matthews B.F., Profiling local gene expression changes associated with *Eimeria maxima* and *Eimeria acervulina* using cDNA microarray, Appl. Microbiol. Biotechnol. 62 (2003) 392–399.

[12] Quackenbush J., Microarray data normalization and transformation, Nat. Genet. 32 (Suppl.) (2002) 496–501.

[13] Rathod P.K., Ganesan K., Hayward R.E., Bozdech Z., DeRisi J.L., DNA microarrays for malaria, Trends Parasitol. 18 (2002) 39–45.

[14] Sørensen P., Bonnet A., Buitenhuis B., Closset R., Déjean S., Delmas C., Duval M., Glass L., Hedegaard J., Hornshoj H., Hulsegge I., Jaffrézic F., Jensen K., Jiang L., de Koning D.J., Lê Cao K.-A., Nie H., Petzl W., Pool M.H., Robert-Granié C., San Cristobal M., Lund M.S., van Schotshorst E.M., Schuberth H.-J., Seyfert H.-M., Tosser-Klopp G., Waddington D., Watson M., Yang W., Zerbe H., Analysis of the real EADGENE data set: Multivariate approaches and post analysis, Genet. Sel. Evol. 39 (2007) 651–668.

[15] Watson M., Pérez-Alegre M., Baron M.D., Delmas C., Dovč P., Duval M., Foulley J.-L., Garrido-Pavón J.J., Hulsegge I., Jaffrézic F., Jiménez-Marín Á., Lavrič M., Lê Cao K.-A., Marot G., Mouzaki D., Pool M.H., Robert-Granié C., San Cristobal M., Tosser-Klopp G., Waddington D., de Koning D.J., Analysis of a simulated microarray dataset: Comparison of methods for data normalisation and detection of differential expression., Genet. Sel. Evol. 39 (2007) 669–683.

Genet. Sel. Evol. 39 (2007) 633–650
© INRA, EDP Sciences, 2007
DOI: 10.1051/gse:2007029

Available online at:
www.gse-journal.org

Original article

Analysis of the real EADGENE data set: Comparison of methods and guidelines for data normalisation and selection of differentially expressed genes (*Open Access publication*)

Florence JAFFRÉZIC[a*], Dirk-Jan DE KONING[b], Paul J. BOETTCHER[c], Agnès BONNET[d], Bart BUITENHUIS[e], Rodrigue CLOSSET[f], Sébastien DÉJEAN[g], Céline DELMAS[h], Johanne C. DETILLEUX[i], Peter DOVČ[j], Mylène DUVAL[h], Jean-Louis FOULLEY[a], Jakob HEDEGAARD[e], Henrik HORNSHØJ[e], Ina HULSEGGE[k], Luc JANSS[e], Kirsty JENSEN[b], Li JIANG[e], Miha LAVRIČ[j], Kim-Anh LÊ CAO[g,h], Mogens Sandø LUND[e], Roberto MALINVERNI[c], Guillemette MAROT[a], Haisheng NIE[l], Wolfram PETZL[m], Marco H. POOL[k], Christèle ROBERT-GRANIÉ[h], Magali SAN CRISTOBAL[d], Evert M. VAN SCHOTHORST[n], Hans-Joachim SCHUBERTH[o], Peter SØRENSEN[e], Alessandra STELLA[c], Gwenola TOSSER-KLOPP[d], David WADDINGTON[b], Michael WATSON[p], Wei YANG[q], Holm ZERBE[m], Hans-Martin SEYFERT[q]

[a] INRA, UR337, Jouy-en-Josas, France (INRA_J); [b] Roslin Institute, Roslin, UK (ROSLIN); [c] Parco Tecnologico Padano, Lodi, Italy (PTP); [d] INRA, UMR444, Castanet-Tolosan, France (INRA_T); [e] University of Aarhus, Tjele, Denmark (AARHUS); [f] University of Liège, Liège, Belgium (ULg2); [g] Université Paul Sabatier, Toulouse, France (INRA_T); [h] INRA, UR631, Castanet-Tolosan, France (INRA_T); [i] Faculty of Veterinary Medicine, University of Liège, Liège, Belgium (ULg1); [j] University of Ljubljana, Slovenia (SLN); [k] Animal Sciences Group Wageningen UR, Lelystad, The Netherlands; [l] Wageningen University and Research Centre, Wageningen, The Netherlands (WUR); [m] Ludwig-Maximilians-University, Munich, Germany; [n] RIKILT-Institute of Food Safety, Wageningen, The Netherlands (WUR); [o] University of Veterinary Medicine, Hannover, Germany; [p] Institute for Animal Health, Compton, UK (IAH); [q] Research Institute for the Biology of Farm Animals, Dummerstorf, Germany

(Received 10 May 2007; accepted 6 July 2007)

* Corresponding author: florence.jaffrezic@jouy.inra.fr

Abstract – A large variety of methods has been proposed in the literature for microarray data analysis. The aim of this paper was to present techniques used by the EADGENE (European Animal Disease Genomics Network of Excellence) WP1.4 participants for data quality control, normalisation and statistical methods for the detection of differentially expressed genes in order to provide some more general data analysis guidelines. All the workshop participants were given a real data set obtained in an EADGENE funded microarray study looking at the gene expression changes following artificial infection with two different mastitis causing bacteria: *Escherichia coli* and *Staphylococcus aureus*. It was reassuring to see that most of the teams found the same main biological results. In fact, most of the differentially expressed genes were found for infection by *E. coli* between uninfected and 24 h challenged udder quarters. Very little transcriptional variation was observed for the bacteria *S. aureus*. Lists of differentially expressed genes found by the different research teams were, however, quite dependent on the method used, especially concerning the data quality control step. These analyses also emphasised a biological problem of cross-talk between infected and uninfected quarters which will have to be dealt with for further microarray studies.

quality control / differentially expressed genes / mastitis resistance / microarray data / normalisation

1. INTRODUCTION

Microarray analyses have been highlighted as an area of high priority within the European Animal Disease Genomics Network of Excellence (EADGENE), to study host-pathogen interactions in animals. Microarrays give the possibility to study the changes of expression of thousands of genes simultaneously depending on the pathogen.

A large variety of methods for normalising and analysing microarray data has, however, been proposed in the literature, and there is still no clear consensus about which analysis process is recommended. The aim of this joint research work was to review the methods and software packages used by the EADGENE partners and to provide some general guidelines for further analyses. To achieve this goal, a real data set was distributed among the workshop participants. The real data was provided by an EADGENE funded microarray study looking at the gene expression changes following artificial infection of cows with two different mastitis causing bacteria: *Escherichia coli* and *Staphylococcus aureus*. The effect of artificial infection was tested over time in 12 dairy cows using three udder quarters in each cow for different time points following infection and one for the control sample. The study included two species of bacteria as well as several time-points, resulting in a true analytical challenge (48 microarrays in total). The EADGENE partners who provided the data were RIBFA and the Roslin Institute.

In this paper three main steps of microarray data analysis will be discussed: data quality control, normalisation and statistical methods for the detection of differentially expressed genes. For each of these steps, the techniques used by the workshop participants will be presented and compared.

2. MATERIALS AND METHODS

2.1. Presentation of the data

2.1.1. Comparison of E. coli vs. S. aureus *elicited mastitis in cows using transcriptomic profiling*

The outcome of an udder infection (mastitis) is influenced by the species of the infecting bacteria. Coliform bacteria, *e.g. E. coli*, tend to cause acute infections with severe inflammatory symptoms, while others, like *S. aureus* often result in chronic infections with less severe symptoms. The molecular causes underpinning these differences in host pathogen interactions are largely unknown. Here, we established a strictly controlled animal model to allow for a systematic analysis of the different immune responses elicited by *E. coli vs. S. aureus*, using strains of both pathogen species previously isolated from field cases of mastitis. Healthy heifers were infected in the fourth month of their first lactation. None of the cows had suffered a previous udder infection and their somatic cell counts were well below 100 000 cells per mL of milk.

Three trials were conducted, each comprising four animals. First, 500 CFU of our asseverated *E. coli* strain 1303 were infected into udder quarters at time 0, 12 and 18 h. The fourth quarter was kept as a control. The animals were culled after 24 h and sampled. All animals showed signs of acute clinical mastitis by 12 h after challenge: increased somatic cell count (SCC), decreased milk yield, leucopenia, fever and udder swelling. Quantitative RT-PCR analysis revealed that the expression of Toll-like receptor (TLR) 2, TLR4 and beta-defensin-encoding genes was greatly enhanced in the 24 h infected quarters, while the relative mRNA copy numbers remained low in the uninfected control quarters, which is coherent with the microarray results presented below. Secondly, animals infected with 10 000 CFU of the *S. aureus* strain 1027 in a similar scheme over 24 h (n = 4) showed no or only modest clinical signs of mastitis. No evidence of alteration in TLR or beta-defensin-encoding indicator genes for activated innate immune defense was found. In the third trial, four animals were infected with the *S. aureus* pathogen. For each of them (i) two quarters were infected at time 0, (ii) a third quarter at time 60 h, and animals

were killed after 72 h. Hence, there were two quarters per animal with *S. aureus* inoculated for 72 h, one quarter with the pathogen inoculated for 12 h and again one control quarter. *S. aureus* caused clinical symptoms and increased expression of the TLR and beta-defensin-encoding indicator genes in this third group of animals, infected over 72 h (n = 4).

Assignment of the animals to become inoculated with *E. coli* or *S. aureus* was completely at random and arbitrary. The three trials were conducted at three different days. Inter-animal transmission can be excluded, thanks to proper handling of the inoculates. The identity of the pathogens were verified from re-isolates of milk samples. In addition to the classical microbiological verification, strain identity was verified using diagnostic digests of pathogen residential plasmids as criteria.

The clinical and qRT-PCR data proved that the *E. coli* infected animals all developed symptoms of acute mastitis, earlier than 24 h after infection. *S. aureus* pathogens, however, needed more time to elicit not only clear infection related symptoms of mastitis, but also the activation of the immune defense within the udder. We also noted a clear host-individual influence in this regards. Samples from all these udder quarters were carefully asseverated and stored in liquid nitrogen, for subsequent DNA-microarray analyses.

The microarray experiment was carried out using the Bovine 20K array (ARK-Genomics). A common reference design was used and the reference sample was made up of all 48 RNA samples. The reference sample was labelled with Cy3 and the treatment with Cy5 on each microarray slide. All samples were collected in Hannover (Germany) by Holm Zerbe, Hans-Joachim Schuberth, and Wolfram Petzl, and had been validated by Hans-Martin Seyfert in Dummerstorf (Germany). The samples were shipped to the Roslin Institute for transcriptome profiling by Elizabeth Glass and Kirsty Jensen.

The Bovine 20K microarray was subdivided in 48 blocks, with 12 rows and 4 columns. Each of the 48 resulting blocks was printed with its own unique print-tip (*i.e.* there are 48 print-tips). Each block consisted of 30 sub-grid rows and 30 sub-grid columns. Almost all (19 705) features were printed in duplicate within the same block, 324 printed 4 times and 2 printed 12 times. Annotations were provided by Mark Fell of the Roslin Institute and were distributed among the workshop participants. The microarrays were scanned and data were extracted using Bluefuse (http://www.cambridgebluegnome.com/bluefuse.htm). Bluefuse does not provide an estimate of the background intensity, and therefore no further background correction was possible on these data.

2.2. Normalisation of the data

2.2.1. Data quality control

Several quality control procedures were used by the authors and Table I presents an overview of these techniques. Most of the teams used the spot quality indicators provided by the scanning software (Bluefuse) to make decisions about excluding spots from the analysis. There are several indicators of quality provided by the Bluefuse software: (a) the probability that a clone is expressed in the tissue studied (PON) with a value between 0 and 1; (b) a manual quality flag from A (good) to E (bad); (c) a compound 'confidence' quality indicator between 0 and 1; and (d) a binary quality indicator that is 0 (bad) or 1 (good). The simplest approaches were to remove spots with manual flags or with Bluefuse flag values equal to D or E because their confidence levels were lower than 0.30 (meaning a poor quality of spot). In more sophisticated approaches, raw data were visualised using R-LimmaGUI [15] to check the overall quality by several criteria, such as M boxplots, M-A plots, and Cy5-Cy3 scatter plots. INRA_T pointed out, using simple descriptive statistics that array BTK2-74 was different from the other slides given the mean, minimum and maximum, and should be deleted from the analysis. M-A plots of the raw data were atypical and showed a clear 'fishtail' pattern for low intensity spots, where the log-ratios (M) diverged, as shown in Figure 1. This indicated relatively noisy data due to many spots with low intensities. ROSLIN therefore proposed to add 2^8 to all the channel intensities. IDL deleted spots with intensities above 65 000 (oversaturated spots) or with values within the experimental error, *i.e.* spots smaller than 400 [8]. AARHUS suggested a quality weighting of the data [9] by down-weighting the spots with low quality based on Bluefuse 'Confidence' or 'P ON' measurements. For all teams, data were log2 transformed and the log-ratio between Cy5 and the reference Cy3 was considered as the observed intensity.

2.2.2. Correction for spatial and intensity-dependent bias

Normalisation of the data is a two-step process including first a correction for spatial bias, and second a correction for intensity-dependent bias. Correction for spatial bias was usually carried out separately for each block (print-tip) of each array, by either subtracting the median for each block, or by subtracting the corresponding row and column means (RC correction, excluding control spots) [1]. The intensity dependent bias was removed by either block-Loess correction [14], or by a global Loess correction [17]. Two levels for each of

Table I. Overview of the methods used by each team for data quality control and normalisation.

Team	Quality Control (QC)	Normalisation	Softwares
ROSLIN	(1) Genespring. (2) 2^8 added to all spot intensities. Three plots per slide: MA, print-tip box-plot and spatial plot. Summary statistics were used.	Four normalisations: global and local dye and spatial correction.	Genespring Bioconductor (Limma and Marray) with changes.
AARHUS	Diagnostic plots. Weights for data quality were used based on 'Confidence' and 'P-ON' weights.	Print-tip Loess correction was used with different weights.	Bioconductor Limma.
INRA_T	Clones with quality control flag A, B or C (n = 6948) were kept. Slide BTK2-74 was omitted.	Arrays and genes were mean centered.	R.
IDL	Would use background information but it was not provided here with Bluefuse. Spot edits, oversaturated spots and spots smaller than 400 were removed.	Print tip Loess. Heterogeneous variance correction [8].	Limma http://www.asgbioinformatics.wur.nl
ULg1		2-step Wolfinger procedure. Random effect model with a fixed dye effect, a random print-tip effect and an interaction term.	SAS®
SLN		Samples were normalised to uninfected group.	(1) Genespring. (2) Orange http://www.ailab.si/orange
IAH	Blank, auto_excl and man_excl spots were removed. Replicate spots were averaged.	Median normalisation for each slide. Scale normalisation between arrays.	Co-Express programmed in R.
INRA_J	Manual flags were considered as missing (man_excl).	Global Loess and print-tip correction. Replicate spots were considered as independent observations.	R.
ULg2	'DNA', 'blank', 'buffer', 'nothing', 'light reference' were deleted. Spots were deleted if quality is bad (E) for 25 consecutive spots.	Global Loess.	R, Maanova.
WUR	M box-plots across slides. Slide BTK2-74 was omitted. Background (BG) was estimated and spots where Signal/BG <2 were deleted (from 20 k to 8.7 k genes).	Global Loess. 'Rikilt' normalisation.	Genemaths XT Limma.
PTP	Quality measures were used to delete the worst spots. Non-relevant spots were also deleted.	Global Loess, Mixed model across slides (Wolfinger 2-step procedure [16]).	SAS®

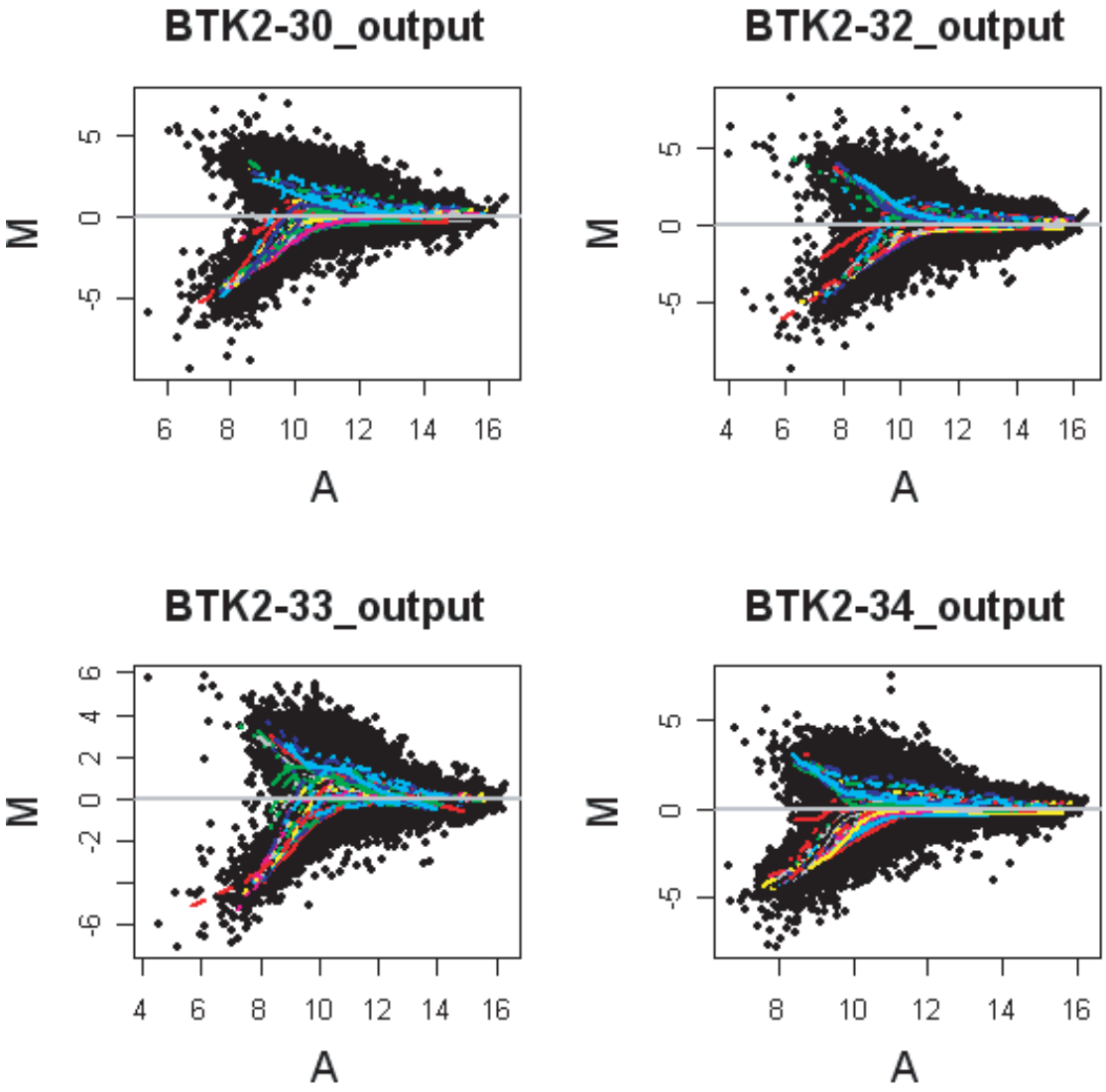


Figure 1. The "fishtail" appearance of M-A plots for the raw data for slides 1–4. Lines are Loess curves for each of the 48 print-tips. Control spots were omitted.

the two normalisation steps were examined by ROSLIN to check whether these steps should be global (*i.e.*, chip-wide) or local (*i.e.*, print-tip). The choice was informed by comparisons of summary measures of M-A plots, spatial heat diagrams and print-tip box-plots for the raw data and all four normalisations. The local spatial bias (RC correction) and local intensity-bias (MA normalisation) were found here to perform consistently well regarding the spatial plot in the F-test of differences between blocks in M values, the M-A plot in the F-test of a block MA correction *versus* a chip-wide MA correction, and the print-tip box-plots in the mean inter-quartile range of M. This local RC-MA normalisation

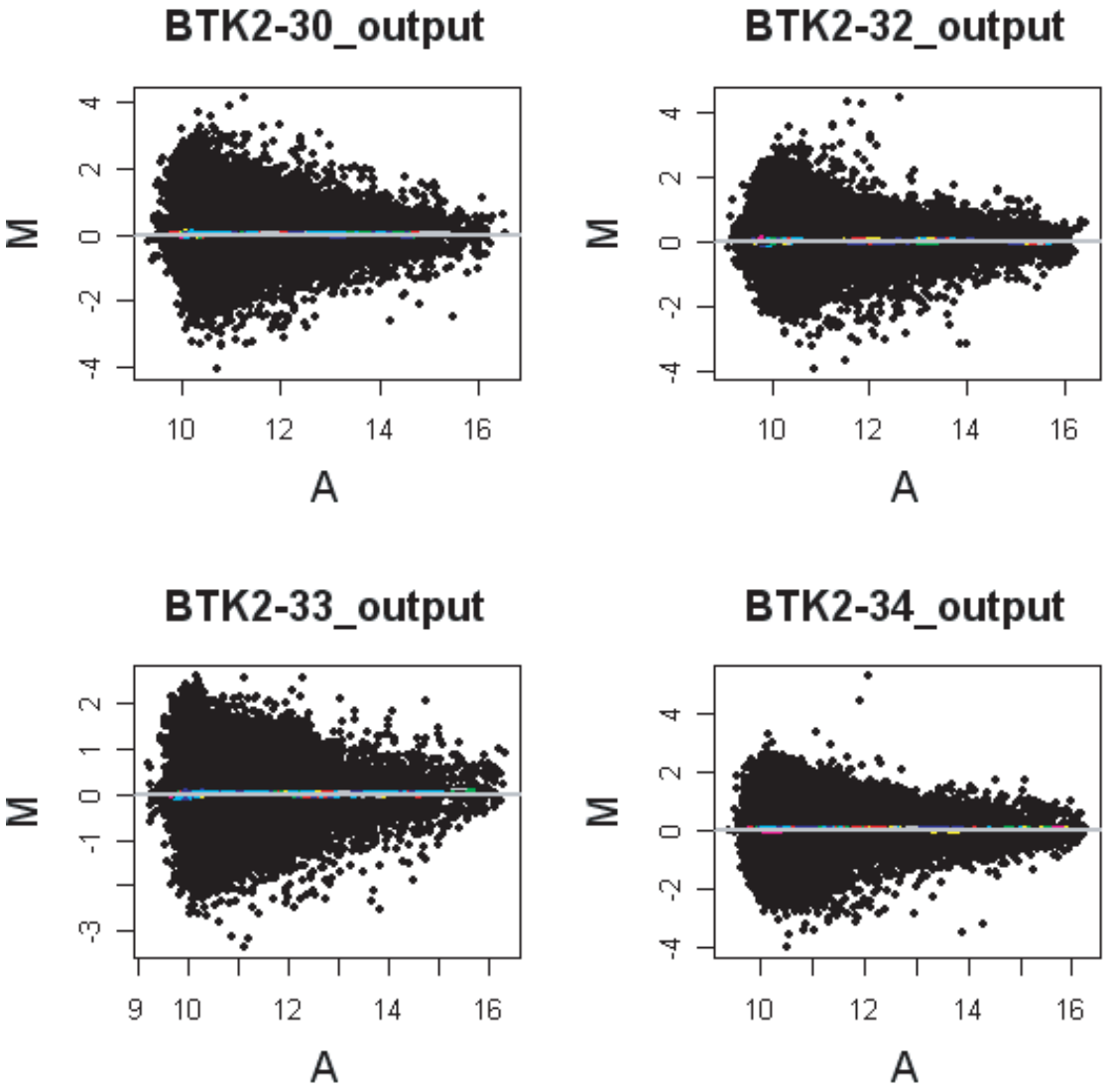


Figure 2. M-A plots after normalisation for ROSLIN team with a print-tip Loess correction for slides 1 to 4.

was therefore chosen by this team for normalising the data. Figure 2 shows the corresponding M-A plots after normalisation. Since *E. coli* and *S. aureus* samples were hybridised always at the same channel and against a common reference, the setup of this experiment requires no dye swap effect correction, which is often a source of experimental noise.

Another possible approach for data normalisation is to use an ANOVA model. This approach was used here by two teams (ULg1 and PTP) using a two-stage mixed-model approach [5] with the Proc Mixed SAS® procedure. In the first stage, initial models were fitted to each array separately to take account of the experimental systematic effects on the base-2 logarithm of the

pixel values. The model included a fixed dye effect, a random print-tip effect and an interaction between dye and print-tip effects. Effects of print-tip were considered as random because of the manufacturing variation expected between print-tips. Residuals obtained from this model were then analysed to find differentially expressed genes.

It has to be emphasised that all the genes were used in the normalisation procedures presented above, based on the underlying assumption that most of the genes are not differentially expressed and that the observed differences are only due to technical artefacts. This assumption has to be checked for every experiment and may sometimes not be verified, especially when using dedicated chips.

2.2.3. Software packages used for the data normalisation

Four teams (ROSLIN, AARHUS, IDL, WUR) used the Bioconductor package Limma – Linear Models for Microarray Analysis [13] in R for data normalisation. A bioinformatics pipeline was developed by IDL to handle both data normalisation and detection of differentially expressed genes accessible at http://www.ASGbioinformatics.wur.nl. The SAS® software was used for normalisation using an ANOVA model.

2.3. Methods used to find differentially expressed genes

Three main biological questions were investigated on this data set: which genes are differentially expressed (1) between the two types of infection (*E. coli* and *S. aureus*); (2) over time within each bacteria; and (3) across time and bacteria. Table II presents an overview of the statistical methods used by each team to find differentially expressed genes.

2.3.1. ANOVA approach with different variance models

Three teams (ROSLIN, AARHUS, IDL) used for this part of the analyses the Bioconductor R package Limma [13], which allows complex designs and provides robust t- and F-statistics for differential gene expression by the use of empirical Bayes methods (eBayes) for shrinking the residual variances of genes towards their approximate median value. This approach is based on an inverse chi-square prior on the variances [12]. The linear model used here accounted for within-array replicate spots and included the effects of time and

Table II. Overview of the methods and software packages used by each team for the detection of differentially expressed genes.

Team	Comparison	Single gene analysis Statistical method	Softwares
ROSLIN	Uninfected quarters, between infections and across time within each infection.	(1) Genespring. (2) Limma + FDR.	(1) Genespring. (2) Bioconductor Limma with eBayes correction.
AARHUS	Infected at different times *vs.* uninfected for each separate experiment.	Limma + FDR.	Bioconductor Limma with eBayes correction.
INRA_T	*E. coli* infected *vs.* unifected.	Fisher test with FDR.	R.
IDL	For both *E. coli* and *S. aureus*: infected at different times *vs.* uninfected.	Limma FDR + Fold change cut-off.	Limma https://www.asgbioinformatics.wur.nl
ULg1	Infected *vs.* uninfected. *E. coli vs. S. aureus*	SAS®, Wolfinger mixed model, FDR correction.	SAS®
SLN	Infected *vs.* uninfected.	2-fold change. Anova with FDR.	(1) Genespring. (2) Orange http://www.ailab.si/orange
IAH	*S. aureus vs. E. coli*	Derived from clusters and differences over time.	Co-Express programmed in R.
INRA_J	Different time points within infection or same time point between infections.	Anova model, Structural mixed model for variances, Time course with EDGE.	'SMVar' function programmed in R EDGE [3].
ULg2	For *S. aureus*: infected *vs.* uninfected.	Bayesian analysis of variance.	R, BAMarray http://www.bamarray.com
WUR	Infection and time combined.	2-way ANOVA. 1-way ANOVA.	TIGR [11].
PTP	Within and between infections at different times.	Linear mixed model: bacteria effect, and time as a categorical or continuous variable.	SAS®.

challenging bacteria. Differentially expressed genes between types of infection were tested based on the robust t-statistics and differential expression of genes over the different time points used a moderated F-test. Another approach also based on an ANOVA model but with a different variance model was used by INRA_J. It is based on a structural mixed model on the variances [7] and is implemented in R in the 'SMVar' function. Here, a fixed condition effect and a random gene effect were considered to model the log of the variances. Two other teams (WUR and ULg2) used TIGR Multiple Experimental Viewer v4.0 [11] and the BAMarray software [6] for Bayesian analysis of variance, respectively. In the latter approach, genes are clustered into groups of equal variances and data are rescaled to satisfy the equal variance assumption. Then, a hierarchical Bayesian model is used to synthesise information across all genes simultaneously, and estimated effects for genes unlikely to be differentially expressed are shrinked to zero to enhance patterns of interest.

2.3.2. Models for time-course study

In the ANOVA models presented above, observations were assumed to be independent, which was not the case in this time-course study since measurements were made at different time points for each animal. Three longitudinal approaches were proposed here to take these correlations into account and find differentially expressed genes over time in the two infections.

The first approach was performed by PTP using the Mixed procedure of SAS®. A gene-by-gene analysis was performed on the residuals obtained from the normalisation process. The effects included in this linear mixed model were the following: a fixed bacteria effect, a non-parametric mean curve by fitting the time effect as a qualitative variable or a parametric function of time (linear and quadratic regression on time), and the interactions between bacteria and these time effects. A linear random regression model was considered (using a random cow effect and an interaction with time). A quadratic random regression model was also investigated but did not converge. For each gene-specific model, custom hypothesis tests were constructed to determine whether gene expression was different between healthy and infected quarters, or between quarters infected with *E. coli* and *S. aureus* at different times.

The second longitudinal approach considered in this workshop by INRA_J was based on the Edge package [3]. In this gene-specific model, the population average time curve was modelled using a natural cubic spline function and the correlation structure was fitted with a random intercept. Two biological questions can be addressed with this approach. First, is the effect for each gene

constant for each infection. Second, is the expression pattern over time, *i.e.* the average time curve, for each gene the same in the two infections.

In the last approach, an ANOVA was performed by INRA_T on the expression value for each *E. coli* clone, with the time factor as an explanatory variable (4 levels: 0, 6, 12 and 24 h). A standard Fisher test was used to test the effect of time on each gene. After selection of the differentially expressed genes over time, a clustering approach based on smoothed expression curves [4,10] was used to find clusters of genes with similar expression profiles. This second step is presented in the post-analysis paper.

2.3.3. Correction for multiple tests

Regarding the correction for multiple tests, all teams used the classical Benjamini and Hochberg [2] correction at a 5% False Discovery Rate (FDR) threshold, either using R functions or the SAS® Multitest procedure.

3. RESULTS

Although various methods were applied for normalising and analysing the data, it was reassuring to see that most of the teams found similar biological results. First, it was found that the largest number of differentially expressed genes was obtained when comparing samples from udder tissue challenged for 24 h with *E. coli* to non-challenged tissue. In contrast, challenging with *S. aureus* did not result in a dramatic transcriptional response. Second, quite a large number of differentially expressed genes were detected at time zero between the two groups of infections. This showed a cross-talk between udder quarters or an invasion by immune cells from the infected quarters, since all udders were collected simultaneously at the end of exposure. We will present here the results obtained with the Bioconductor Limma package (ROSLIN in Tab. II) which was used by many teams and was shown to perform well for differential gene expression analysis.

The uninfected quarters from the *E. coli* infection exhibited differential expression in 402 clones representing 359 genes compared to the *S. aureus* uninfected quarters. The most up-regulated genes included metallothioneins and lipopolysaccharide binding protein, indicating that an immune response has been triggered in the uninfected quarters. Furthermore, the MHC class II invariant chain molecule, CD74, was down-regulated, suggesting that the cell populations present in the mammary gland quarter had altered in response to the infection of neighbouring quarters. Considerable overlap was observed in

the gene lists from the *E. coli* uninfected quarters and the 331 genes declared to be differentially expressed at 6 h post *E. coli* infection. More than 600 clones, representing 538 genes exhibited differential expression at 12 h post infection, and the number of differentially expressed genes reached a maximum at 24 h post infection when the transcription of 1190 genes was affected. Many of the most up-regulated genes at this time point are associated with the influx of neutrophils into the infected gland, including S100 calcium binding proteins A8, A9 and A12, colony stimulating factor 3 and several chemoattractants for neutrophils, *e.g.* interleukin 8 and chemokine (C-X-C motif) ligand 1 and 2.

No gene was identified as being significantly differentially expressed during *S. aureus* infection using the cut-off FDR value of 0.05. This lack of statistical support principally results from high levels of variation between the biological replicates. This may be an artefact of the experimental procedure; large mammary gland samples were collected for RNA extraction which may have comprised variable amounts of *S. aureus* infected tissue, because the bacteria causes a localised infection. Therefore the gene lists were expanded to include those genes with t-test p-values less than or equal to 0.01 and a fold change greater than or equal to 1.5. At 6 h post *S. aureus* infection, 154 genes exhibited differential expression. The most highly up-regulated gene was lactotransferrin, an antimicrobial protein secreted in milk. Interestingly, this gene was only observed to be up-regulated at 24 h post *E. coli* infection. At 12 and 24 h post infection 182 and 266 genes were declared differentially expressed, respectively. However, the number decreased to 97 by 72 h post infection. There was some overlap between the lists of differentially expressed genes during *E. coli* and *S. aureus* infections, including the up-regulation of superoxide dismutase and the down-regulation of interleukin 7. The analysis of the microarray data identified two putative genes that may be indicative of *S. aureus* infection. Leucine rich repeat kinase 2 was down-regulated at all 4 time points during *S. aureus* infection but not during *E. coli* infection. In addition, a clone (AJ814901) whose sequence currently only matches EST was also down-regulated during *S. aureus* infection.

Various comparisons of lists of differentially expressed genes found by the EADGENE teams were performed. We focussed mainly on the comparison of the differentially expressed genes found between time 0 and 24 h for the *E. coli* infection, which exhibited the largest transcriptional response. It was found that although all the teams found a large number of differentially expressed genes between these two time points, the lists of genes were still dependent on the method chosen. Figure 3 gives the Venn diagram for the differentially expressed genes found by three of the teams. They used different data quality

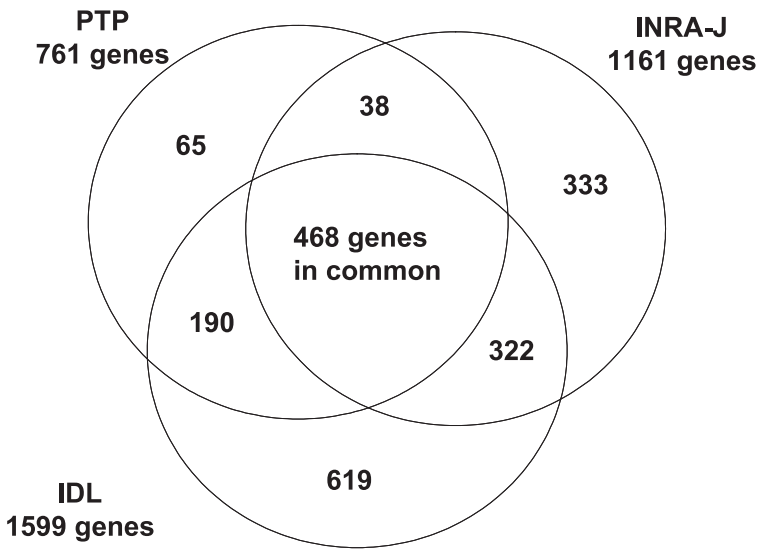


Figure 3. Venn diagram for the lists of differentially expressed genes found for *E. coli* between times 0 and 24 h after infection at a 5% Benjamini and Hochberg threshold for IDL, INRA_J and PTP teams. Normalisation and analyses methods used by these teams are presented in Tables I and II.

control procedures and different normalisation and analyses methods: the IDL team used a print-tip Loess normalisation and the Limma package, INRA_J used a global Loess normalisation and the structural model for variances, and PTP used the global Loess and a 2-step mixed model approach with SAS®. It was found that 468 genes were detected in common for these three teams, and 790 genes were detected in common for IDL and INRA_J. It is interesting to also note that IDL and PTP teams, despite using very different approaches, found 658 genes in common among the 761 genes detected by PTP. When focussing only on the 500 most differentially expressed genes found by the three teams, only 206 genes were found in common for the three approaches, as shown in Figure 4. A larger consistency in the ranking of the genes could have been expected, especially between IDL and INRA_J which used Limma and SMVar shrinkage approaches, respectively. In fact, both teams found here only 272 genes in common among the first 500, although the two methods were found in previous studies [7] to provide very similar results in the detection of differentially expressed genes. The main difference in the analyses performed by these two teams was in the data quality control step. On the contrary, 323 were found in common between IDL and PTP teams, who used very different statistical approaches for the detection of differentially expressed genes but a more similar approach for data quality control, with the removal of oversaturated and low quality spots. The data quality control step therefore appears here

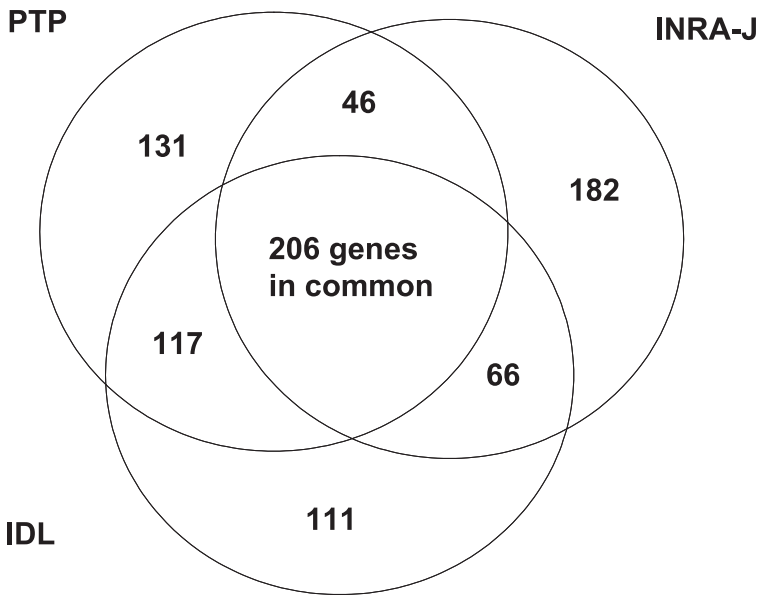


Figure 4. Venn diagram for the 500 first differentially expressed genes found for *E. coli* between times 0 and 24 h after infection for IDL, INRA_J and PTP teams.

to be essential for microarray data analysis but a consensus for best practice still has to be reached.

It has to be emphasised that this study presented the number of genes that were declared to be differentially expressed by statistical methods, at a 5% Benjamini-Hochberg threshold. A biological validation still has to be performed, however, for most of these genes to be able to differentiate between the true positives and the false positives.

4. DISCUSSION

Quality control of the data proved, in this workshop, to be an important first step. Simple summary statistics for each slide, as well as print-tip box-plots, MA and spatial plots can be used for quality checks. Recommendation would be to delete spots that were flagged as low quality, or to perform some quality weighting [9]. It was found here that one of the slides was of very poor quality and had to be deleted from the analysis. Bluefuse does not use a measure of background intensity from pixels around the spot, nor does it make an explicit estimate from within-spot pixels, and therefore the background intensities were not provided.

Following quality control, the normalisation step is divided into two steps: a correction for spatial bias and a correction for intensity-dependent bias. The first step is performed on each block (print-tip) for each array separately either

by subtracting the median for each block, subtracting the corresponding row and column means (excluding control spots) or by including a random array block effect in the Wolfinger *et al.* [16] two-step mixed model approach. The second step is performed using either a local or global Loess correction. Using various quality control plots, one of the teams found that the local Loess correction was the most adapted for this data set. Many teams used the Bioconductor R package Limma to perform this normalisation. More diversity was observed among the teams for the data quality control step than for the normalisation step.

Two main approaches were used for the statistical analysis of these data. The first approach was based on ANOVA models and allowed the detection of differentially expressed genes using two by two comparisons with robust t-tests. The construction of these robust t-statistics was based either on the eBayes Limma shrinkage [13] or on a structural mixed model on variances [7] – SMVar function in R. These analyses provided lists of genes that were differentially expressed within each infection at different time points, as well as between the two infections. The second approach was based on longitudinal models and took into account the correlations between measurements involved in this time-course study. For this second set of analyses, a random regression model was used with SAS® in a two-step mixed model approach, and the EDGE package developed by Dabney *et al.* [3] was applied to these data. These analyses allowed the detection of genes that had a pattern of expression changing over time or that differed for both infections. Since these data come from a longitudinal study, it is advisable here to use the latter approaches that take into account correlation between measurements rather than the ANOVA based models which assume independence of the observations. Correction for multiple tests was performed by all the teams using Benjamini and Hochberg [2] FDR approach at a 5% threshold.

Although various quality control procedures, data normalisation and analysis methods were used, all the teams generally obtained the same main biological results. In fact, all participants found that most of the differentially expressed genes were found between the uninfected group and quarters that had been challenged by the *E. coli* pathogens for 24 h. On the contrary, very little transcriptional response was observed for the *S. aureus* infection.

It can be argued, however, that the robustness observed here concerning the biological results may be due to the extremely large transcriptional response with the *E. coli* infection. These conclusions may therefore not be generalised to other experiments with only small transcriptional changes. Moreover, several methods used here such as the shrinkage approaches (Limma, SMVar,

BAMarray) were designed to improve the performance under high-noise, low replicate, small-change settings. The *E. coli* data, which exhibited a very large transcriptional response, may therefore not allow pointing out the subtle differences between these various methods.

All the teams pointed out the heterogeneity between the two uninfected groups which should have been comparable but exhibited an unexpectedly large number of differentially expressed genes. This observation raised an important biological and experimental design problem about cross-talking between udder quarters. This issue will be studied more thoroughly by the EADGENE biologists in further experiments.

A comparison of the lists of differentially expressed genes found by the workshop participants was performed for *E. coli* between times 0 and 24 h. Due to the various methods used for normalising and analysing the data, the lists were not exactly similar. It was reassuring, however, to find that even using two very different approaches, (1) normalisation by print-tip Loess and analysis with Limma in R; and (2) global Loess and Wolfinger's two-step mixed model approach in SAS®, the lists of differentially expressed genes still remained quite similar.

Here all participants had the same raw data set to analyse. Comparison of methods may have been easier, however, if each step had been evaluated separately: first, data quality control, then normalisation on a common previously cleaned data set and finally detection of differentially expressed genes on a common previously normalised set of data. Criteria to compare procedures for data quality control is still an open and essential issue for microarray data analysis.

ACKNOWLEDGEMENTS

The authors wish to acknowledge Caroline Channing, Karin Smedegard and WP1.4 for organising the workshop, Zerbe *et al.* for providing the real data sets and EADGENE for financial support (EU Contract No. FOOD-CT-2004-506416).

REFERENCES

[1] Baird D., Johnstone P., Wilson T., Normalization of microarray data using a spatial mixed model analysis which includes splines, Bioinformatics 20 (2004) 3196–3205.

[2] Benjamini Y., Hochberg Y., Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. B 85 (1995) 289–300.

[3] Dabney A.R., Leek J.T., Monsen E., Storey J.D., Edge manual, Department of Biostatistics, University of Washington, http://faculty.washington.edu/jstorey/edge, 2006.
[4] Déjean S., Martin P.G.P., Baccini A., Besse P., Clustering time series gene expression data using smoothing spline derivatives, EURASIP J. Bioinformatics Syst. Biol. (2007) ID 70561.
[5] Gibson G., Wolfinger R.D., Gene expression profiling using mixed models, in: Saxton A.M. (Ed.), Genetic analysis of complex trait using SAS®, SAS® User Press, Cary NC, 2004, Chap. 11, pp. 251–278.
[6] Ishwaran H., Rao J.S., Kogalur U.B., BAMarray™: Java software for Bayesian analysis of variance for microarray data, BMC Bioinformatics 7 (2006) 59.
[7] Jaffrézic F., Marot G., Degrelle S., Hue I., Foulley J.-L., A structural mixed model for variances in differential gene expression studies, Genet. Res. 89 (2007) 19–25.
[8] Pool M.H., Hulsegge B., Janss L.L.G., Background bias on cDNA micro-arrays, EAAP, Uppsala, Sweden, 2005.
[9] Ritchie M.E., Diyagama D., Neilson J., van Laar R., Dobrovic A., Holloway A., Smyth G.K., Empirical array quality weights for microarray data, BMC Bioinformatics 7 (2006) 261, http://www.biomedcentral.com/1471-2105/7/261
[10] Robert-Granié C., Baccini A., Besse P., Déjean S., Ferré P.J., Liaubet L., Martin P.G.P., San Cristobal M., Kinetics analysis of microarray data using semiparametric mixed models, 8th World Congress on Genetics Applied to Livestock Production, Belo-Horizonte, Brazil, August 13–18, 2006.
[11] Saeed A.I., Sharov V., White J., Li J., Liang W., Bhagabati N., Braisted J., Klapa M., Currier T., Thiagarajan M., Sturn A., Snuffin M., Rezantsev A., Popov D., Ryltsov A., Kostukovich E., Borisovsky I., Liu Z., Vinsavich A., Trush V., Quackenbush J., TM4: a free, open-source system for microarray data management and analysis, Biotechniques 34 (2003) 374–378.
[12] Smyth G.K., Linear models and empirical Bayes methods for assessing differential expression in microarray experiments, Statist. Appl. Genet. Mol. Biol. 3 (2004) 3.
[13] Smyth G.K., Limma: linear models for microarray data, in: Gentleman R.C., Carey V.J., Dudoit S., Irizarry R., Huber W. (Eds.), Bioinformatics and Computional Biology using R and Biocondutor, Springer, New York, 2005, pp. 397–420.
[14] Smyth G.K., Speed T., Normalization of cDNA microarray data, Methods 31 (2003) 265–273.
[15] Wettenhall J.M., Smyth G.K., LimmaGUI: A graphical user interface for linear modeling of microarray data, Bioinformatics 20 (2004) 3705–3706.
[16] Wolfinger R.D., Gibson G., Wolfinger E.D., Bennett L., Hamadeh H., Bushel P., Afshari C., Paules R.S., Assessing gene significance from cDNA microarray expression data via mixed models, J. Comp. Biol. 8 (2001) 625–637.
[17] Yang Y., Dudoit S., Luu P., Peng V., Ngai J., Speed T., Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, Nucleic Acids Res. 30 (2002) e15.

Genet. Sel. Evol. 39 (2007) 669–683
© INRA, EDP Sciences, 2007
DOI: 10.1051/gse:2007031

Available online at:
www.gse-journal.org

Original article

Analysis of a simulated microarray dataset: Comparison of methods for data normalisation and detection of differential expression (*Open Access publication*)

Michael WATSON[a*], Mónica PÉREZ-ALEGRE[b], Michael Denis BARON[c], Céline DELMAS[d], Peter DOVČ[e], Mylène DUVAL[d], Jean-Louis FOULLEY[f], Juan José GARRIDO-PAVÓN[b], Ina HULSEGGE[g], Florence JAFFRÉZIC[f], Ángeles JIMÉNEZ-MARÍN[b], Miha LAVRIČ[e], Kim-Anh LÊ CAO[h], Guillemette MAROT[f], Daphné MOUZAKI[h], Marco H. POOL[c], Christèle ROBERT-GRANIÉ[d], Magali SAN CRISTOBAL[d], Gwenola TOSSER-KLOPP[d], David WADDINGTON[h], Dirk-Jan DE KONING[h]

[a] Institute for Animal Health, Compton, UK (IAH_C)
[b] University of Cordoba, Cordoba, Spain (CDB)
[c] Institute for Animal Health, Pirbright, UK (IAH_P)
[d] INRA, Castanet-Tolosan, France (INRA_T)
[e] University of Ljubljana, Slovenia (SLN)
[f] INRA, Jouy-en-Josas, France (INRA_J)
[g] Animal Sciences Group Wageningen UR, Lelystad, NL (IDL)
[h] Roslin Institute, Roslin, UK (ROSLIN)

(Received 10 May 2007; accepted 10 July 2007)

Abstract – Microarrays allow researchers to measure the expression of thousands of genes in a single experiment. Before statistical comparisons can be made, the data must be assessed for quality and normalisation procedures must be applied, of which many have been proposed. Methods of comparing the normalised data are also abundant, and no clear consensus has yet been reached. The purpose of this paper was to compare those methods used by the EADGENE network on a very noisy simulated data set. With the *a priori* knowledge of which genes are differentially expressed, it is possible to compare the success of each approach quantitatively. Use of an intensity-dependent normalisation procedure was common, as was correction for

[*] Corresponding author: michael.watson@bbsrc.ac.uk
Institute for Animal Health Informatic groups, Compton Laboratory, Compton RG20 7 NN Newbury Bershive, UK.

multiple testing. Most variety in performance resulted from differing approaches to data quality and the use of different statistical tests. Very few of the methods used any kind of background correction. A number of approaches achieved a success rate of 95% or above, with relatively small numbers of false positives and negatives. Applying stringent spot selection criteria and elimination of data did not improve the false positive rate and greatly increased the false negative rate. However, most approaches performed well, and it is encouraging that widely available techniques can achieve such good results on a very noisy data set.

gene expression / two colour microarray / simulation / statistical analysis

1. INTRODUCTION

Microarrays have become a standard tool for the exploration of global gene expression changes at the cellular level, allowing researchers to measure the expression of thousands of genes in a single experiment [16]. The hypothesis underlying the approach is that the measured intensity for each gene on the array is proportional to its relative expression. Thus, biologically relevant differences, changes and patterns may be elucidated by applying statistical methods to compare different biological states for each gene. However, before comparisons can be made, a number of normalisation steps should be taken in order to remove systematic errors and ensure the gene expression measurements are comparable across arrays [15]. There is no clear consensus in the community about which methods to use, though several reviews have been published [8, 12]. After normalisation and statistical tests have been applied, there is an additional problem of multiple testing. Due to the high number of tests taking place (many thousands in most cases), the resulting P-values must be adjusted in order to control or estimate the error rate (see [14] for a review).

The aim of this paper was to summarise and compare the many methods used throughout the EADGENE network (http://www.eadgene.org) for microarray analysis, and compare the results, with the final aim of producing a guide for best practice within the network [4]. This paper describes a variety of methods applied to a simulated data set produced by the SIMAGE package [1]. The data set is a simple comparison of two biological states on ten arrays, with dye-balance. A number of data quality, normalisation and analysis steps were used in various combinations, with differing results.

1.1. The data

SIMAGE takes a number of parameters, which were produced using a slide from the real data set as an example [4]. The input values that were used for the current simulations are given in Table I. The simulated data consists of

ten microarrays each of which represent a direct comparison between different biological samples from situation A and B with a dye balance. SIMAGE assumes a common variance for all genes, something which may not be true for real data. Each slide had 2400 genes in duplicate, with 48 blocks arranged in 12 rows and 4 columns (100 spots per block). Each block was "printed" with a unique print tip. In the simulated data 624 genes were differentially expressed: 264 were up-regulated from A to B while 360 were down regulated. This information was only provided to the participants at the end of the workshop. The simulated data are available upon request from D.J. de Koning (DJ.dekoning@bbsrc.ac.uk).

The data are very noisy with high levels of technical bias and thus provided a serious challenge for the various analysis methods that were applied. Many spots reported background higher than foreground, and others reported a zero foreground signal. Image plots of the arrays showed clear spatial biases in both foreground and background intensities (Fig. 1). Spots, scratches and stripes of background variation are clearly visible, which have been simulated using the "hair" and "disc" parameters of SIMAGE.

All of the slides show a clear relationship between M (log ratio) and A (average log intensity), and the plots in Figure 2 are exemplars. Slides 3, 5, 6, 7, 9 and 10 displayed a negative relationship between M and A, whilst the others displayed a positive relationship. Slides 6 and 9 showed an obvious non-linear relationship between M and A, but only slide 2 levels off with higher values of A. Finally, Figure 3 shows the range of M values for each array under three different normalisation strategies: none (Fig. 3a), LOESS (Fig. 3b) and LOESS followed by scale normalisation between arrays (Fig. 3c) [17, 19]. It can be seen that before normalisation there is a clear difference in both the median log ratios and the range of log ratios across slides.

This data set was subject to a total of 12 different analysis methods, encompassing a variety of techniques for assessing data quality, normalisation and detecting differential expression. These methods are described in detail and the results of each presented and compared. The results are then discussed in relation to the best methods to use for analysing extremely noisy microarray data.

2. MATERIALS AND METHODS

2.1. Preprocessing and normalisation procedures

A variety of pre-processing and normalisation procedures were used in combination with the twelve different methods, and these are summarised in

Table I. Settings for Simage simulation software.

Array number of grid rows	12
Array number of grid columns	4
Number of spots in a grid row	10
Number of spots in a grid column	10
Number of spot pins	48
Number of technical replicates	2
Number of genes	0
Number of slides	10
Perform dye swaps	yes
Gene expression filter	yes
Reset gene filter for each slide	no
Mean signal	10.33
Change in $\log_2$ ratio due to upregulation	1.07
Change in $\log_2$ ratio due to downregulation	–1.26
Variance of gene expression	2.7
% of upregulated genes	15
% of downregulated genes	11
Correlation between channels	1
Dye filter	yes
Reset dye filter for each slide	yes
Channel variation	0.2
Gene × Dye	0
Error filter	yes
Reset error filter for each slide	yes
Random noise standard deviation	0.62
Tail behaviour in the MA plot	0.108
Non-linearity filter	yes
Reset non-linearity filter for each slide	yes
Non-linearity parameter curvature	0.2
Non-linearity parameter tilt	4.5
Non-linearity from scanner filter	yes
Reset non-linearity scanner filter for each slide	yes
Scanning device bias	0.04
Spotpin deviation filter	yes
Reset spotpin filter for each slide	no
Spotpin variation	0.32
Background filter	yes
Reset background filter for each slide	yes
Number of background densities	5
Mean standard deviation per background density	0.2
Maximum of the background signal relative to the non-background signals	50
Standard deviation of the random noise for the background signals	0.1
Background gradient filter	no
Reset gradient filter for each slide	yes
Maximum slope of the linear tilt	700
Missing values filter	yes
Reset missing spots filter for each slide	yes
Number of hairs	3
Maximum length of hair	20
Number of discs	4
Average radius disc	10
Number of missing spots	50

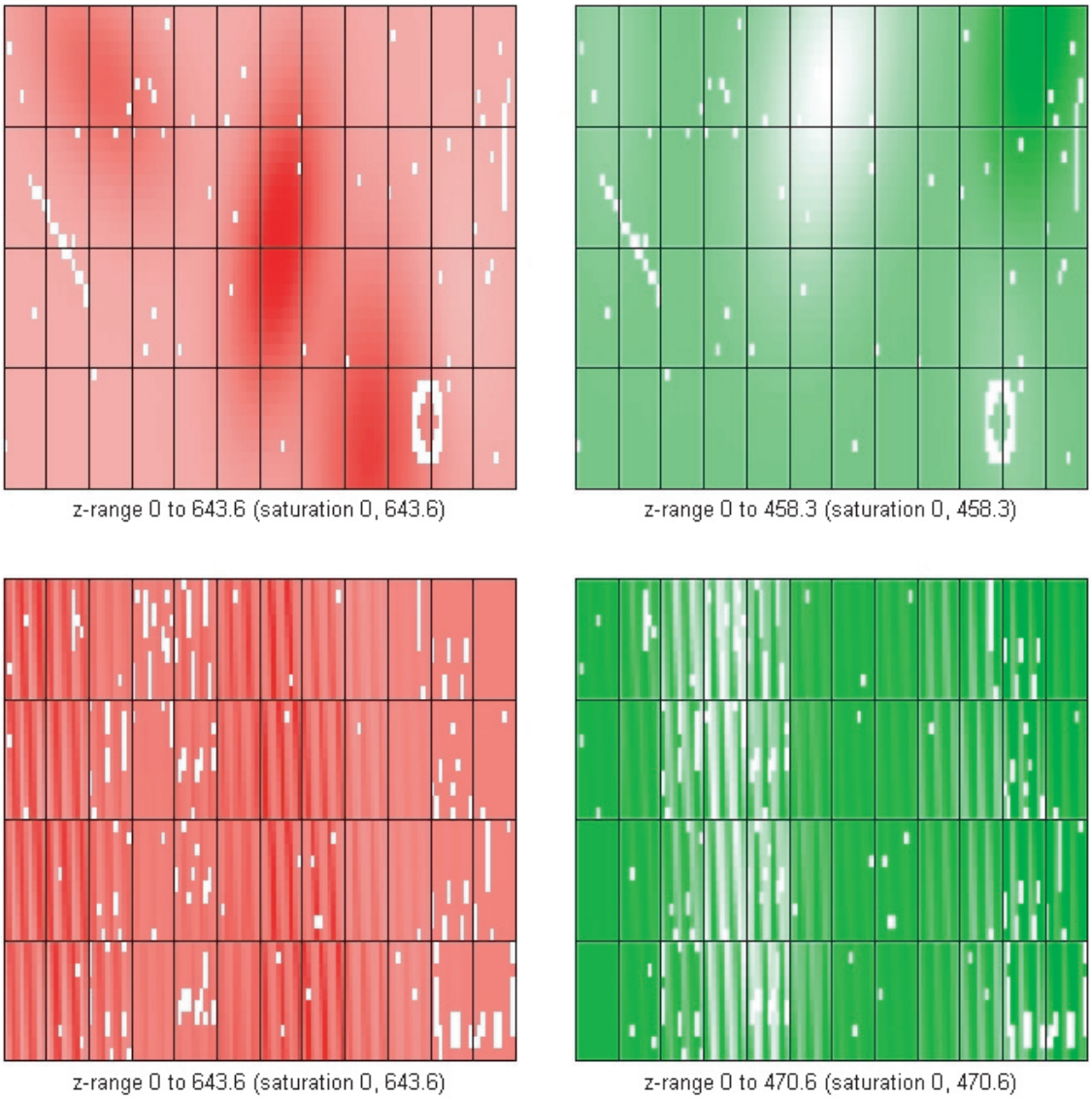


Figure 1. Example background plots. The top two images show the background for Cy5 and Cy3 in slide 9, and the bottom two images show the same for slide 10.

Table II. Only one method, IDL1, chose to perform background correction. Some methods chose to eliminate spots, or give them zero weighting, depending on particular data quality statistics; these included having foreground less than a given multiple of background, saturated spots and spots whose intensity was zero. IAH_P1 and IDL1 also removed entire slides considered to have poor data quality. Both IAH_P and IDL submitted two approaches, one based on strict quality control and normalisation, and the second less strict.

Most approaches applied a version of LOWESS or LOESS normalisation, either globally or per print-tip [19]. This is in recognition of the clear relationship between M and A. Only ROSLIN (assessed normalisation by row and

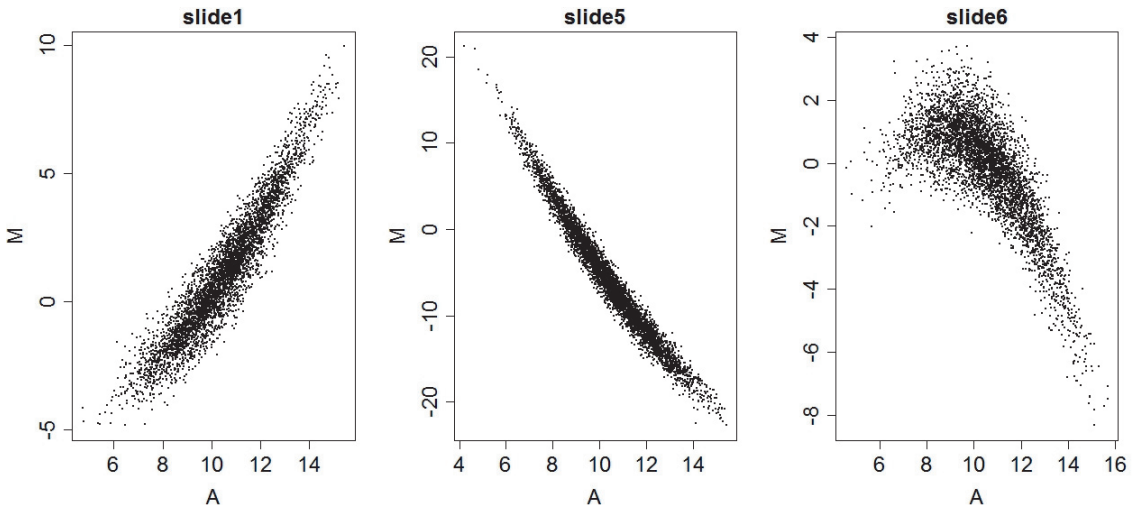


Figure 2. MA-plots of slides 1, 5 and 6. These slides are examples of the three patterns displayed by the simulated data in the MA-space: positive correlation, negative correlation and a more pronounced non-linear correlation.

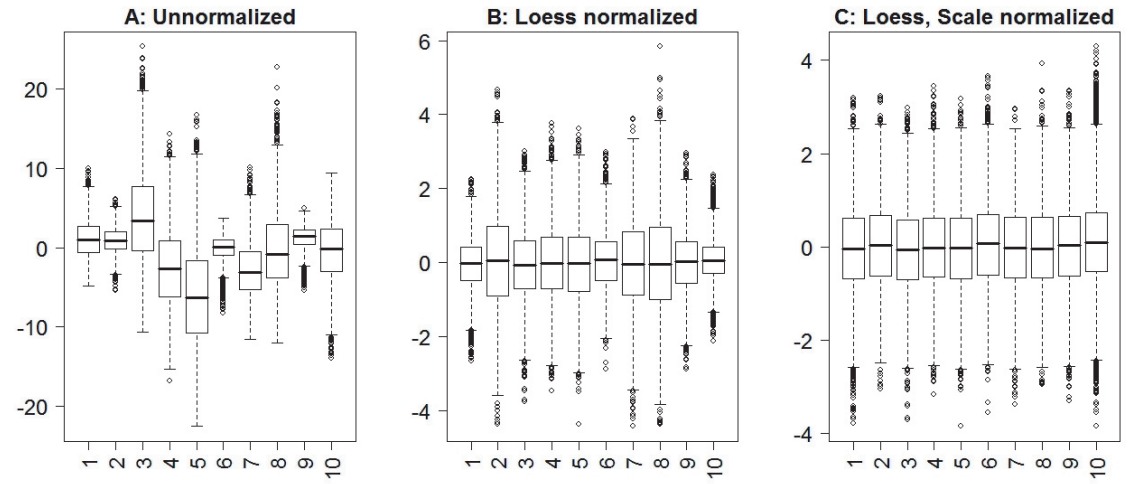


Figure 3. Boxplots of M values ($\log_2$(cy5/cy3)) across the 10 arrays for three normalisation strategies: (A) Unnormalised data, (B) LOESS normalised data, and (C) LOESS followed by scale normalised data.

column and found not needed) and INRA_J (correction by block) applied any further spatial normalisation. SLN1 and SLN2 applied median normalisation. Finally, only IDL attempted any correction between arrays by fitting a monotonic spline in MA-space to correct for heterogeneous variance. The smoothing function was fitted to the absolute log ratios (M-values) across the log mean intensities (A-values), and corrected for. This ensured that the variance in M values was consistent across arrays.

Table II. Summary of the 12 methods used for analysing the simulated data. "Analysis name" is the name of the analysis method, "Data quality procedures" describe the methods approach to data quality, "Background correction" whether background correction was carried out, "Normalisation" describes the normalisation method and "Differential expression" describes the method's approach to finding differentially expressed genes.

Analysis name	Data quality procedures	Background correction	Normalisation	Differential expression
IAH_P1	Eliminated spots with net intensity < 0. Slides 5, 6 and 9 deleted	No	global LOWESS	Limma; FDR correction
IAH_P2	Slides 5, 6 and 9 deleted	No	global LOWESS	Limma; FDR correction
IDL1	Eliminated • control spots • null spots • oversaturated spots • values < 3* SD bgnd. Slides 5 and 7 deleted	Yes	printtip LOWESS; monotonic spline correction	Limma; FDR correction
IDL2		No	global LOWESS; monotonic spline correction	Limma; FDR correction
INRA_J	Spots == zero removed	No	LOWESS; median normalisation by block	structural mixed model; FDR correction
INRA_T1	Spots == zero removed	No	global LOWESS	Student statistic; FDR correction
INRA_T2	Spots == zero removed	No	global LOWESS	Student statistic; Duval correction
INRA_T3	Spots == zero removed	No	global LOWESS	Student statistic; Bordes correction
ROSLIN	Spots == zero removed	No	printtip LOWESS; row-column normalisation	Limma; FDR correction
SLN2	Only use data where FG > 1.5* BG	No	median normalisation	Anova (Orange)
CDB	Elimination of spots with huge M-values	No	printtip LOWESS	fold change cut-off (+/–0.9)
SLN1	Excluded BG > FG	No	median normalisation	Anova (GeneSpring)

2.2. Methods for finding differentially expressed genes

Table II summarises the twelve methods used for analysing the simulated data set. Most variation in the methods came from the area of quality control, with different groups excluding different genes/arrays based on a wide variety of criteria, and correction for multiple testing.

Almost all analysis methods used some variation of linear modelling followed by correction for multiple testing to find differentially expressed genes. The most common of those used was the limma package, which adjusts the t-statistics by empirical Bayes shrinkage of the residual standard errors toward a common value (near to the median) [17]. IAH_P and ROSLIN fitted

a coefficient for the dye-effect for each gene, which was found to be non-significant. IAH_P also adjusted the default estimated proportion of differentially regulated genes in the eBayes procedure to 0.2 once it became clear that a high percentage of the genes in the dataset were differentially regulated. This ensured a good estimate of the posterior probability of differential expression.

Of those that did not use limma, both SLV and SLN2 used an ANOVA approach, implemented in GeneSpring [9] and Orange [5] respectively. INRA_J used a structural mixed model, more completely described in Jaffrézic *et al.* [11]. CDB employed a cut-off value for the mean log ratio to define the proportion of differentially expressed genes [10, 18]. INRA_T presented three methods all based on a classic Student statistic and an empirical variance calculated for each gene, but with the P-values adjusted according to Benjamini and Hochberg [2], Duval *et al.* (partial sums of ordered t-statistics) [6, 7] and Bordes *et al.* (mixture of central and non-central t-statistics) [3]. Apart from INRA_T, those methods that corrected P-values for multiple testing did so using the FDR as described by Benjamini and Hochberg [2]. All corrections for multiple testing were carried out at the 5% level.

All methods treated the 10 arrays as separate, biological replicates apart from ROSLIN, who treated the dye-swaps as technical replicates. The INRA_J and the three INRA_T methods treated replicate spots as independent measures, resulting in up to 20 values per gene, whereas the other methods averaged over replicate spots. INRA_T reported that preliminary analysis showed very few differences between treating duplicates as independent or by averaging them.

3. RESULTS

Table III summarises the results for the analysis of the simulated data set. In terms of the total number of errors made (false positives + false negatives), methods INRA_T2 and INRA_T3 excelled with only 17 and 12 errors respectively. In terms of the least number of false negatives, methods IDL2 and INRA_T1 performed best, having both missed only one gene that was differentially expressed. Many of the analysis methods scored upwards of 95% correctly identified genes. Of those that did not, IAH_P1 and IDL1 operated strict quality control measures, and may have eliminated a number of differentially expressed genes from the analysis. When the number of correct genes is expressed as a percentage of the number of genes each method identified, these methods too show greater than 95% correctly identified genes. Those methods based on traditional statistics performed less well than those methods

Table III. Summary of the results of the analysis of the simulated data set. Table shows the number of genes identified by each method as differentially expressed, the number correct, the number of false positives and negatives, the number of correctly identified genes as a % of the total number of differentially expressed genes (624) and as a % of the number of genes identified for each method.

Analysis	No	Correct	False +	False –	Correct/total	Correct/identified
IAH_P1	499	485	14	139	77.72	97.19
IAH_P2	608	592	16	32	94.87	97.37
IDL1	304	289	15	335	46.31	95.07
IDL2	642	623	19	1	99.84	97.04
INRA_J	663	614	49	10	98.40	92.61
INRA_T1	649	623	26	1	99.84	95.99
INRA_T2	629	618	11	6	99.04	98.25
INRA_T3	622	617	5	7	98.88	99.20
ROSLIN	628	600	28	24	96.15	95.54
SLN2	171	128	43	496	20.51	74.85
CDB	67	44	23	580	7.05	65.67
SLN1	3	3	0	621	0.48	100.00

specifically designed with microarray data in mind. CDB chose a fold-change cut-off above which genes were flagged as significant, set at a $\log_2$ ratio of +/– 0.9. SLN1 analysed the dye-swap slides separately, which will have reduced the statistical power of the analysis, combining the results afterwards. This resulted in only three genes identified as differentially expressed; however, all were correct. SLN2 identified 171 genes as differentially expressed, but also showed a relatively high number of false positives and negatives.

Table IV shows the top ten differentially expressed genes that were missed by the 12 methods (false negatives). One gene, gene 203, was missed by every analysis method. Genes 2221 and 465 were missed by all but two methods, those being IDL2 and INRA_T1 in both cases. These genes are characterised by log ratios that do not necessarily match their direction of regulation and very large standard deviations relative to the normalised mean log ratios.

Table V shows the top ten genes wrongly identified as differentially expressed by the 12 analysis methods (false positives). Gene 1819 was identified as differentially expressed in 8 of the 12 methods; however, given that CDB, SLN1 and SLN2 identified very few genes in total, this means that only one of the more accurate methods correctly called this gene as **not** differentially expressed, and that is INRA_T3. Moving further down, there are four genes called as false positives in six of the methods, though there is no consistency

Table IV. The top ten genes identified as false negatives in the 12 analysis methods. Table contains the gene id (gene), mean and standard deviation of the unnormalised log ratio (M and SD), mean and standard deviation of the LOESS normalised log ratio (M LOESS and SD LOESS), the number of methods in which the gene was a false negative (Count) and the direction of regulation from SIMAGE (Regulated).

Gene	M	SD	M LOESS	SD LOESS	Count	Regulated
gene203	–1.35	3.25	–0.01	0.65	12	up
gene2221	–1.71	3.14	–0.40	0.39	10	up
gene465	–0.70	3.00	–0.39	0.59	10	up
gene1411	2.74	6.80	–0.48	0.67	9	up
gene352	0.63	3.97	–0.39	0.84	8	up
gene1448	–4.24	6.26	–1.32	1.87	7	down
gene1580	–2.12	3.58	–0.58	0.89	7	up
gene1667	2.59	6.61	0.69	0.78	7	up
gene1704	–2.26	4.16	–0.46	1.11	7	up
gene90	3.06	6.53	–0.47	1.01	7	up

Table V. The top ten genes identified as false positives in the 12 analysis methods. The table contains the gene id (gene), mean and standard deviation of the unnormalised log ratio (M and SD), mean and standard deviation of the LOESS normalised log ratio (M LOESS and SD LOESS) and the number of methods in which the gene was a false positive (Count).

Gene	M	SD	M LOESS	SD LOESS	Count
gene1819	1.93	4.67	0.50	0.42	8
gene2262	–0.65	3.45	0.65	0.67	6
gene555	0.72	3.75	–0.55	0.65	6
gene995	0.18	2.93	0.60	0.65	6
gene999	–0.18	3.30	0.54	0.38	6
gene1258	1.98	5.04	0.48	0.52	5
gene1324	–0.12	3.34	0.60	0.44	5
gene1654	0.33	3.69	0.52	0.61	4
gene2069	–0.35	4.04	–0.33	0.51	4
gene2110	3.40	5.07	0.49	0.61	4

shown in which methods identified those four correctly or incorrectly. These genes are characterised by standard deviations that are about equal to the normalised log ratios, in contrast to the false negatives.

4. DISCUSSION

After the comparison, we are in the unique position of knowing *a priori* which and how many genes were differentially expressed, however before starting the analysis none of the groups had the information and only a very noisy data set was provided. Each group applied a different variety of techniques to find the differentially expressed genes. In some cases, the data were put into a standardised pipeline, and in others the analysis was customised to this data set.

It is interesting to note that only one method used any kind of background subtraction. This was due to researchers recognising that although some slides displayed high background, there was little relationship with spot foreground, and therefore subtracting background would have removed many spots from the analysis with no resulting benefit. A consensus in the wider community on background correction has yet to be reached, however the partners within the EADGENE network appeared to have done so, with all but one partner deciding not to correct for local background when analysing this data set.

Applying stringent spot quality procedures and subsequent elimination of both spots and slides from the analysis, as seen in IAH_P1 and IDL1, did not greatly lessen the number of false positives, but greatly increased the number of false negatives. The increase in false negatives was much larger than the corresponding decrease in false positives. This suggests that, when dealing with noisy data, care must be taken to eliminate only data for which a real physical source of error can be identified, *e.g.* detector saturation during scanning. In the case of the data analysed here some of the simulated backgrounds were high, leading some groups to reject those spots; in fact, rejecting the estimated backgrounds was the best approach, since eliminating data from the analysis leads to the elimination of significantly differentially expressed genes with no associated benefit.

It is clear from the relationship between M and A that an intensity dependent normalisation should be used on these data and most groups reflected that by choosing to use LOWESS/LOESS normalisation. The spatial biases shown in the background suggest that perhaps a spatial normalisation technique should be used, yet only two investigated the need for it: INRA_J and ROSLIN. The differences seen in the range of raw log ratios between slides

suggest that a between-slides normalisation method would have been appropriate, yet only IDL attempted to do so. Figure 3 shows the range of M values for each array under three different normalisation strategies: none (Fig. 3a), LOESS (Fig. 3b) and LOESS followed by scale normalisation between arrays (Fig. 3c) [17, 19]. Figure 3a shows that there is a large amount of variation in the range of M values between slides, and Figure 3b shows that that variation is not entirely removed by LOESS normalisation alone. Figure 3c shows the most uniform distribution of M values across arrays, as can be expected given the normalisation strategy. Whether or not this is desirable depends on the context of the experiment. For example, one would expect technical replicates to have very similar distributions, whereas biological replicates may not. In this experiment, if we assume that the dye-swapped arrays are technical replicates, then array pairs 5 and 6, and 9 and 10, represent technical replicates of one another, yet show vastly differing ranges of M values (Fig. 3a), adding weight to the argument for between array normalisation. The failure to apply additional normalisation steps after the first may have been due to fears of "over-fitting" the data. However, ROSLIN report that additional analyses were carried out on the data with between-slides variation-standardisation applied, and an additional 23 genes were identified, 12 of which were differentially expressed, the other 11 being false positives (data not shown).

The approaches may be split into traditional and more sophisticated methods of analysis. SLN1, SLN2 and CDB employed more traditional methods (analysis of variance and fold-change cut-off), whereas the others employed methods shown to be of particular use with microarray data. The authors from CDB wish it to be known that theirs was only a preliminary analysis. DNMAD [18] and GEPAS [10] are sophisticated tools for the analysis of microarray data, and it is unfortunate that some of their more sophisticated methods were not brought to bear on the simulated data. The more traditional methods were also more conservative, identifying fewer genes in total as differentially regulated. They did not, however, have correspondingly smaller false positive rates.

Examination of the genes consistently appearing as false negatives or false positives reveals predictable trends. Consistent false negatives showed very high variation about the mean, whereas consistent false positives showed much less. The simulation software, SIMAGE, gives the same ratio to all genes designated up- or down-regulated, therefore any difference between genes designated as up- or down-regulated is solely down to noise modelled by the software. Those genes consistently identified as false negatives simply received more noise, and those consistently identified as false positives received less.

Overall, given that this was a noisy data set, it is promising that such high numbers of correctly identified genes can be achieved. The trade off between false positives and false negatives can clearly be seen and suggests that elimination of data due to poor spot quality measures does not pay off in terms of the decrease in false positives given the large increase in false negatives. Correction for the false discovery rate (FDR) [2] was the most commonly used technique for adjusting P-values. However, a direct comparison of multiple testing procedures occurred in the INRA_T analyses, with the two novel methods presented out-performing the FDR procedure proposed by Benjamini and Hochberg [2] in terms of error rate; the mixture model described by Bordes *et al.* [3] performed particularly well. The performance of the INRA_T methods is of note given that similar gene-by-gene methods have been shown to lack power in comparison to shrinkage methods such as limma [17] and the structural model [11]. It may be that the data was sufficiently well replicated to overcome this. In addition, this data set has been simulated with homogeneous variances, and this assumption may not hold true for real data sets.

It should be noted that the simulated data represents a well replicated experiment, with ten replicates for a single comparison. This no doubt lends a great deal of power to the analyses. Additional power was achieved by INRA_J and the three INRA_T methods by treating replicate spots as independent measures, resulting in up to twenty measurements per gene. Although these four techniques showed very good results, comparable results were achieved by ROSLIN, IAH_P2 and IDL2, showing that the increase in replication from ten to twenty did not greatly improve the results. In fact, the IAH_P2 analysis, which eliminated 3 out of the 10 slides but still achieved very high success rates, showed that this data set was probably over-endowed with replicates, beyond what would normally be found in a real experiment. Repeating the analyses with a smaller number of replicates may be informative. Kooperberg *et al.* [13] compared methods for analysing microarray experiments with small numbers of replicates and concluded that the best methods were those which took an empirical Bayes approach (*e.g.* [17], used in some analyses presented here) and those that combined similar experiments.

ACKNOWLEDGEMENTS

The authors acknowledge the Danish participants and WP1.4 for organising the workshop and EADGENE for financial support (EU Contract No. FOOD-CT-2004-506416).

REFERENCES

[1] Albers C.J., Jansen R.C., Kok J., Kuipers O.P., van Hijum S.A., SIMAGE: simulation of DNA-microarray gene expression data, BMC Bioinformatics 7 (2006) 205.

[2] Benjamini Y., Hochberg Y., Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. Royal Stat. Soc. Ser. B 57 (1995) 289–300.

[3] Bordes L., Delmas C., Vandekerkhove P., Semiparametric estimation of a two component mixture model when a component is known, Scand. J. Stat. 33 (2006) 733–752.

[4] de Koning D.J., Jaffrézic F., Lund M.S., Watson M., Channing C., Hulsegge I., Pool M.H., Buitenhuis B., Hedegaard J., Hornshøj H., Jiang L., Sørensen P., Marot G., Delmas C., Lê Cao K.-A., San Cristobal M., Baron M.D., Malinverni R., Stella A., Brunner R.M., Seyfert H.-M., Jensen K., Mouzaki D., Waddington D., Jiménez-Marín Á., Pérez-Alegre M., Pérez-Reinado E., Closset R., Detilleux J.C., Dovč P., Lavrič M., Nie H., Janss L., The EADGENE microarray data analysis workshop, Genet. Sel. Evol. 39 (2007) 621–631.

[5] Demsar J., Zupan B., Leban G., Orange: From Experimental Machine Learning to Interactive Data Mining, White Paper (http://ww.ailab.si/orange) (2004), Faculty of Computer and Information Science, University of Ljubljana.

[6] Duval M., Degrelle S., Delmas C., Hue I., Laurent B., Robert-Granié C., A novel procedure to determine differentially expressed genes between two conditions, 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte (Brazil), August 13–18, 2006.

[7] Duval M., Delmas C., Laurent B., Robert-Granié C., A simple procedure to detect noncentral observations from a sample, http://www.lsp.ups-tlse.fr/Recherche/Publications/2006/duv06.html.

[8] Fujita A., Sato J.R., Rodrigues L. de O., Ferreira C.E., Sogayar M.C., Evaluating different methods of microarray data normalization, BMC Bioinformatics 7 (2006) 469.

[9] GeneSpring GX, http:// www.agilent.com/chem/genespring.

[10] Herrero J., Al-Shahrour F., Díaz-Uriarte R., Mateos A., Vaquerizas J.M., Santoyo J., Dopazo J., GEPAS: A web-based resource for microarray gene expression data analysis, Nucleic Acids Res. 31 (2003) 3461–3467.

[11] Jaffrézic F., Marot G., Degrelle S., Hue I., Foulley J.L., A structural mixed model for variances in differential gene expression studies, Genet. Res. 89 (2007) 19–25.

[12] Jeffery I.B., Higgins D.G., Culhane A.C., Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data, BMC Bioinformatics 7 (2006) 359.

[13] Kooperberg C., Aragaki A., Strand A.D., Olson J.M., Significance testing for small microarray experiments, Stat. Med. 24 (15) (2005) 2281–2298.

[14] Pounds S.B., Estimation and control of multiple testing error rates for microarray studies, Brief. Bioinform. 7 (2006) 25–36.

[15] Quackenbush J., Microarray data normalization and transformation, Nat. Genet. 32 (2002) 496–501.
[16] Schena M., Shalon D., Davis R.W., Brown P.O., Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science 270 (1995) 467–470.
[17] Smyth G.K., Linear models and empirical Bayes methods for assessing differential expression in microarray experiments, Stat. Appl. Genet. Mol. Biol. 3 (2002) Article 3.
[18] Vaquerizas J.M., Dopazo J., Díaz-Uriarte R., DNMAD: web-based diagnosis and normalization for microarray data, Bioinformatics 20 (2002) 3656–3658.
[19] Yang Y.H., Dudoit S., Luu P., Lin D.M., Peng V., Ngai J., Speed T.P., Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, Nucleic Acids Res. 30 (2002) e15.