université
de **BORDEAUX**

Thèse présentée pour obtenir le grade de

DOCTEUR
DE L'UNIVERSITÉ DE BORDEAUX

École doctorale Sociétés, Politique, Santé Publique

Spécialité Santé Publique, option Biostatistique

Thèse préparée dans le cadre du Réseau doctoral en santé publique animé par l'EHESP

Par Boris HEJBLUM

Analyse intégrative de données de grande dimension appliquée à la recherche vaccinale

Integrative analysis of high-dimensional data applied to vaccine research

Sous la direction de Rodolphe THIÉBAUT

co-directeur: François CARON

Soutenue le 6 mars 2015

Membres du jury

RICHARDSON Sylvia	Professor, MRC Biostatistics Unit (Cambridge, UK)	Présidente
BIERNACKI Christophe	Professeur, Université Lille 1 (Lille, France)	Rapporteur
GUIHENNEUC-JOUYAUX Chantal	Professeure, Université Paris Descartes (Paris, France)	Rapporteure
COMMENGES Daniel	Directeur de Recherche, Inserm (Bordeaux, France)	Examinateur
NIKOLSKI Macha	Directeur de Recherche, CNRS (Bordeaux, France)	Examinatrice
CARON François	Research Fellow, University of Oxford (Oxford, UK)	Co-directeur de thèse
THIÉBAUT Rodolphe	Professeur, Université de Bordeaux (Bordeaux, France)	Directeur de thèse

Contents

CONTENTS

Remerciements

Je tiens en premier lieu à remercier **Rodolphe Thiébaut** qui m'a dirigé tout au long de ce travail. Tu as toujours été disponible, et trouvé du temps pour moi dans ton agenda de millionnaire. Tu m'as énormément appris, et je ne cesse d'être impressionné par ta compréhension profonde des différents modèles statistiques. J'espère être un jour aussi expérimenté et clairvoyant que toi. Au cours de ces trois (et quelques) années, ton enthousiasme et ta confiance ont été de puissants moteurs pour ma motivation, et je n'oublierai pas les opportunités que tu m'as offertes. Ce fut un réel plaisir que de travailler avec toi, et j'espère sincèrement que cela pourra continuer.

C'est ensuite tout naturellement que je remercie profondément **François Caron**. Tu m'as adopté en cours de route, et ouvert les portes du monde Bayésien. Tes connaissances, ton sens de la rigueur et ta pédagogie ont été d'une aide précieuse pour produire ce travail et je t'en suis très reconnaissant. Tu as toujours su me remotiver lorsqu'il fallait dériver un produit de Kronecker de plus, tout en me promettant un verre chaleureux dans un de ces vieux pubs d'Oxford. Je suis très fier d'avoir pu travailler sous ta direction et je profite de ces lignes pour t'exprimer mon immense gratitude et reconnaître tout ce que je te dois.

Daniel Commenges, vous m'avez recueilli lors de mon stage de master 2, puis accueilli à bras ouverts dans votre laboratoire. Vous avez été mon premier mentor, en me dirigeant dans un travail de recherche pour lequel, même s'il ne figure pas dans cette thèse, j'éprouve une fierté particulière. Je vous dois beaucoup. Merci.

Chantal Guihenneuc-Jouyaux et **Christophe Biernacki**, merci d'avoir accepté de rapporter cette thèse. Je suis également très honoré de la présence de **Sylvia Richardson** dans mon jury. Je vous en remercie. **Macha Nikolski**, vous avez participé à mon comité de thèse et je suis très heureux de savoir que votre esprit critique veille sur le développement de mes travaux, merci d'avoir une fois de plus pris le temps de m'écouter.

I also would like to thank my immunologist collaborators **Jason Skinner** and **David Furman**. Jason, I had a wonderful time in Dallas and I hope we will see each other again. Je remercie **Lise** et **Damien** que j'ai eu la chance de co-encadrer durant cette thèse. J'espère que ces quelques mois à travailler ensemble vous ont été aussi bénéfiques qu'ils l'ont été pour moi.

Je remercie **Réjane**, **Sandrine** et **Catherine**, pour leur efficacité et leur gentillesse,

sans qui toutes ces conférences et séjours à l'étranger n'auraient pu s'organiser. Je remercie également **Guillaume** qui a toujours répondu à mes demandes plus ou moins techniques. Je remercie aussi l'**équipe de la documentation** qui trouve toujours une solution pour accéder aux articles dont nous avons besoin. Plus largement, je remercie toute l'équipe biostatistique de l'ISPED pour m'avoir fourni un environnement de travail très convivial.

Je remercie aussi particulièrement **Pierre Gay** pour sa gentillesse et sa diligence qui m'ont permis de profiter du formidable **MCIA** (Mésocentre de Calcul Intensif Aquitain) de l'Université de Bordeaux et de l'Université de Pau et des Pays de l'Adour au cours de mes travaux. Je remercie également toute l'équipe de **Qarnot Computing** qui m'a fait profiter de sa belle idée : combiner des calculateurs avec des radiateurs, pour chauffer tout en comptant, à moins que ce ne soit l'inverse. Je remercie le **réseau doctoral de l'École des Hautes Études en Santé Publique** (EHESP) pour m'avoir fait confiance en finançant mon travail durant 3 ans. Je remercie également le professeur **Yves Lévy** ainsi que le **Vaccine Research Institute** pour leur soutien, grâce auquel j'ai notamment pu présenter mon travail dans de nombreuses conférences et nouer des collaborations internationales fructueuses. I am also deeply grateful to all the **participants of the DALIA-1 trial**, as well as to all the **members of the trial committees and of the ANRS/VRI study team** who made it possible for me to work on the DALIA-1 trial data (Trial Committee : Jacques Banchereau, Geneviève Chêne, Carson Harrod, Christine Lacabaratz, Yves Levy, Monica Montes, Karolina Palucka, Laura Richert, Louis Sloan ; DSMB members : William Duncan, Roy M. Gulick, Daniel R. Kuritzkes, James Neaton, Richard Pollard – Chair ; Event Validation Committee members : Joseph Fay, Ronald Mitsuyasu, Jean-Paul Viard ; Trial management : Derek Blankenship, Céline Boucherie, David Jutras, Bryan King, Sophie Pérusat, Elisa Priest, Charlie Quinn, Anna Laura Ross, Mathieu Surenaud ; Vaccine/GMP team : Susan Burkeholder, Amanda Cobb, Charles McWilliams, Jennifer Finholt-Perry, Lee Roberts).

Merci **Robinou** pour tes conseils toujours avisés, ta connivence parisienne, les goûters, ton engouement pour le html, les sessions à la fraîche du samedi matin et tout le reste. Et bien sûr merci à chaque membre du **BéBé club** (**Reto, Paulo, Didi, Dédé, Loulou, Nini** and **Toto**). On rigole bien quand même avec vous les copains. Merci à **Mathieu**, parti voguer vers d'autres horizons mais dont je n'oublierai jamais la passion un peu perverse pour les effets *batch*. Merci à **Laura** pour sa bonne humeur chronique, et enfin merci à ceux qui ont pris la relève dans le grand bureau du 3e, par ordre d'apparition : **Henri** ('Papi'), **Loïc**, et la petite dernière **Perrine**. Chaque matin, je suis pressé d'arriver au bureau pour vous retrouver aux aurores, et chaque soir, c'est toujours à regret que je vous souhaite la bonne aprèm'. Sans oublier l'apparition éclair de la reine de la Kiz', **Chloé**. Tu reviens bientôt et je te laisse Rodolphe quelque temps, profites-en bien. Mélanie, c'est

présentées ici sont appliquées.

Et enfin, la meilleure pour la fin, un immense merci à **Anaïs**, ma chérie, qui illumine chaque jour. Ton amour est une force incroyable. Merci, pour tout.

REMERCIEMENTS

> *Je dédie ce manuscript à mes grand-parents Ginette Buffeteau, Pierre Buffeteau et Samuel Hejblum, dont le plus haut diplôme était le certificat d'étude et qui m'ont transmis le goût du travail bien fait.*

Scientific production

*: equal contribution

Articles

Thesis publications

▶ D. Furman*, B.P. Hejblum*, N. Simon, V. Jojic, C.L. Dekker , R. Thiébaut, R.J. Tibshirani, M.M. Davis, A systems analysis of sex differences reveals an immunosuppressive role for testosterone in the response to influenza vaccination, *Proceedings of the National Academy of Sciences of the United States of America*, 111(2):869–874, 2014.
DOI: 10.1073/pnas.1321060111

▶ B.P. Hejblum, J. Skinner, R. Thiébaut, Time-course Gene Set Analysis for longitudinal gene expression data, *Under revision.*

▶ B.P. Hejblum, F. Caron, R. Thiébaut, Dirichlet process mixture of skew t-distributions for modeling flow cytometry data, *In preperation.*

▶ R. Thiébaut, B. Hejblum, L. Richert, The analysis of "Big Data" in clinical research, *Revue d'Épidémiologie et de Santé Publique*, 62(1):1–4, 2014.
DOI: 10.1016/j.respe.2013.12.021

Communications

Oral communications at international conferences

▶ B. Hejblum, F. Caron, R. Thiébaut, Bayesian analysis of time-course flow cytometry data with Dirichlet process mixture modeling, *27th International Biometric Conference*, Florence, Italy, 2014.

▶ B. Hejblum, R. Genuer, R. Thiébaut, Variable selection in high-dimensional dataset: comparison of sPLS with other approaches in an HIV vaccine trial, *8th International Conference on Partial Least Squares and Related Methods*, Paris, France, 2014.

► R. Thiébaut, B. Hejblum, J. Skinner, M. Montes, G. Chene, K. Palucka, J. Banchereau, Y. Levy, Integrative Analysis of Responses to Dendritic-Cell Vaccination Identifies Signatures Correlated with Control of HIV Replication: The DALIA Trial, *AIDS Vaccine 2013*, Barcelona, Spain, 2013, *AIDS Research and Human Retroviruses* 29 (11), A5-A6.

► B. Hejblum, J.Skinner, R. Thiébaut, Application of Gene Set Analysis of Time-Course gene expression in a HIV vaccine trial, *33rd Annual conference of the International Society for Clinical Biostatistics*, Bergen, Norway, 2012.

Invited talks

► Invited speaker at the Ph.D. students working group of the LSTA (*Laboratoire de Statistique Théorique et Appliquée*) in Paris 6 University, B. Hejblum, F. Caron, R. Thiébaut, Bayesian nonparametric modeling of flow cytometry data with Dirichlet process mixtures, Paris, France, 2014.

► Invited speaker at the first nTaïDA (New Technologies for Autoimmune/Inflammatory Disease Analysis) Workshop "Transcriptome & Exome sequencing" of DHU (*Département Hospitalo-Universitaire*) I2B (Inflammation, Immunopathology & Biotherapy), B. Hejblum, R. Thiébaut, Analysis of repeated measurements of gene expression data: Application to an HIV vaccine, Paris, France, 2015.

Written communications (posters) at international conferences

► B. Hejblum, F. Caron, R. Thiébaut, Hierarchical analysis of time-course flow cytometry data with Dirichlet process mixture modeling, *Medical Research Council Conference on Biostatistics in celebration of the MRC Biostatistics Unit's centenary year*, Cambridge, United Kingdom, 2014.

► B. Hejblum, J. Skinner, R. Thiébaut, Time-course Gene Set Analysis applied in a HIV vaccine trial, *SMPGD 2013: Statistical Methods for (post)-Genomics Data*, Amsterdam, Netherlands, 2013.

Software

► TcGSA: an ® package to analyze longitudinal gene-expression data at the gene set level. Available on CRAN, development version on GitHub.

► NPflow: an ® package to perform clustering of large cell populations with Dirichlet process mixture of skew Normal and skew *t*-distributions. *Under development* .

Notations and abbreviations

Notations

- x: scalar x

- $\boldsymbol{x}$: vector $\boldsymbol{x}$

- $\{\boldsymbol{x}_{a:b}\}$: set of vectors $\{\boldsymbol{x}_a, \boldsymbol{x}_{a+1}, \ldots, \boldsymbol{x}_{b-1}, \boldsymbol{x}_b\}$

- X: matrix X

- X': X transpose

- $p(A)$: probability of A

- $p(A|B)$: probability of A conditional on B

- $\mathbb{E}(X)$: expectation of random variable X

- $X \propto Y$: X is proportional to Y

- $\chi^2_{(k)}$: chi-square distribution with k degrees of freedom

- $\text{Mult}(\boldsymbol{\pi})$: multinomial distribution of probability parameter $\boldsymbol{\pi}$

- $NiW(\boldsymbol{\mu}_0, \kappa_0, \Lambda_0, \lambda_0)$: Normal inverse-Wishart distribution of parameters $(\boldsymbol{\mu}_0, \kappa_0, \Lambda_0, \lambda_0)$

- $sNiW(\boldsymbol{\xi}_0, \boldsymbol{\psi}_0, B_0, \Lambda_0, \lambda_0)$: structured Normal inverse-Wishart distribution of parameters $(\boldsymbol{\xi}_0, \boldsymbol{\psi}_0, B_0, \Lambda_0, \lambda_0)$ [Frühwirth-Schnatter and Pyne, 2010]

- $\Gamma_d(x)$: the d-dimensional gamma function

- $F_d(x)$: the d-dimensional digamma function

Abbreviations

ATI: Antiretroviral Treatment Interruption

ARV: *Antiretroviraux*

cDNA: complementary Deoxyribonucleic Acid

cRNA: complementary Ribonucleic Acid

CRP: Chinese restaurant process

DNA: Deoxyribonucleic Acid

DP: Dirichlet Process

DPM: Dirichlet Process Mixture

EM: Estimation-Maximisation algorithm

FDR: False Discovery Rate

FWER: Family Wise Error Rate

GEM: Griffiths-Engen-McCloskey distribution [Pitman, 2006]

GO: Gene Ontology [Ashburner et al., 2000]

KEGG: Kyoto Encyclopedia of Genes and Genomes [Kanehisa and Goto, 2000]

MAP: Maximum *a posteriori*

MCMC: Monte-Carlo Markov Chain

MLE: Maximum Likelihood Estimation

mRNA: messenger Ribonucleic Acid

MSEP: Mean Squared Error of Prediction

NiW: Normal inverse-Wishart distribution

OLS: Ordinary Least Squares

OR: Odds Ratio

PLS: Partial Least Squares

PRESS: PRediction Error Sum of Squares

RMSEP: Root Mean Squared Error of Prediction

RSS: Residual Sum of Squares

sPLS: sparse Partial Least Squares

sNiW: structured Normal inverse-Wishart distribution [Frühwirth-Schnatter and Pyne, 2010]

TcGSA: Time-course Gene Set Analysis

TIV: Trivalent Inactivated seasonal influenza Vaccine

Thi: Testosterone level high

Tlo Testosterone level low

Résumé substantiel

1   Introduction

L'analyse de données d'expression génique est bien identifiée comme un problème où la grande dimension des données nécessite des outils statistiques spécifiques et sophistiqués. Néanmoins, les mesures phénotypiques effectuées dans le cadre d'essais cliniques vaccinaux sont à l'heure actuelle également devenues des données de grande dimension. Ainsi, la biologie et les études cliniques sont également sujettes à la multiplication des données dans le phénomène actuel représenté par les « Données Massives » (*Big Data*). Par exemple, les populations de cellules immunitaires sont mesurées via des cytomètres à 8 ou 16 couleurs, permettant de distinguer jusqu'à 2^{16} populations de cellules différentes, qui sont elles-mêmes bien souvent suivies sur plusieurs intervalles de temps. Par ailleurs, il existe plusieurs techniques afin d'évaluer la fonction des cellules prélevées : l'ELISPOT, la cytométrie avec marquage intracellulaire et le luminex (quantification de la production de multiples cytokines)... Ces méthodes conduisent à obtenir plusieurs centaines d'observations pour un individu donné, pour un temps de mesure donné. Une analyse intégrative a pour enjeu de relier tous ces différents types de données, données qui sont toutes de grande dimension.

2   Analyse par groupe de gènes de données d'expression génique au cours du temps

Introduction

L'expression génique est un processus dynamique à la racine de tout mécanisme métabolique, qui repose sur les mécanismes de transcription et de traduction de l'ADN. De nombreuses technologies existent à l'heure actuelle pour étudier ce mécanisme, à différents niveaux biologiques. Mais étant de moins en moins chères, les expériences de puces à ADN (*microarray*) sont de plus en plus utilisées pour l'évaluation de l'expression génique au cours du temps. L'analyse de changements temporels de l'expression génique contribue à une meilleure compréhension des mécanismes de régulation des gènes. Plu-

sieurs approches ont été proposées pour analyser ces données longitudinales de grande dimension, gène-par-gène [Storey et al., 2005], en réduisant leur dimension [Liquet et al., 2012], ou bien par groupes de gènes [Wang et al., 2009]. Un groupe de gènes est un ensemble de gènes *a priori* co-régulés ou liés de manière fonctionnelle. Les processus biologiques définis par KEGG ou Gene Ontology, ainsi que les modules fonctionnels définis par Chaussabel et al. [2008] sont des exemples de groupes de gènes. L'analyse par groupes de gènes [Subramanian et al., 2005; Efron and Tibshirani, 2007] est supposée être plus puissante que l'analyse gène-par-gène, car elle peut détecter le changement coordonné de l'expression d'un groupe de gènes sans qu'aucun d'entre eux ne rencontre individuellement un changement très significatif. De plus, le changement de l'ensemble des gènes au sein d'un processus biologique particulier peut avoir plus d'importance sur le plan biologique qu'une modification importante d'un seul gène. Enfin, et à condition que les groupes de gènes soient bien définis, les résultats d'une analyse par groupes de gènes devraient être plus interprétables et plus reproductibles que ceux obtenus par une analyse gène-par-gène [Subramanian et al., 2005].

L'analyse de données d'expression génique répétées au cours du temps par groupes de gènes présente un certain nombre de spécificités. L'une d'entre elles est que les changements dans l'expression génique peuvent être hétérogènes au sein d'un groupe de gènes donné [Efron and Tibshirani, 2007; Ackermann and Strimmer, 2009]. La fréquence de ce phénomène [Ackermann and Strimmer, 2009] nous empêche de l'ignorer, d'autant que l'on ne s'attend pas à ce que les gènes impliqués dans un même processus biologique varient de manière synchrone.

Le choix de l'hypothèse nulle est un des éléments déterminants dans une méthode d'analyse par groupe de gènes. On peut les répartir en deux grandes classes [Goeman and Bühlmann, 2007] : i) les hypothèses nulles compétitives, qui testent les gènes d'un groupe contre tous les gènes en dehors de ce groupe ; ii) les hypothèses autonomes, qui n'utilisent que les gènes à l'intérieur du groupe. Ici nous nous intéressons à un sous-type d'hypothèse autonome, « l'hypothèse nulle mixte » : H_0 : *les gènes à l'intérieur d'un groupe de gènes sont stables au cours du temps*. Cette hypothèse nulle mixte permet de détecter à la fois les groupes de gènes aux changements homogènes, et ceux aux changements hétérogènes. A noter que la modification de l'expression d'un gène au cours du temps pourra ne pas être correctement capturée par la modélisation mais de toute façon sera diagnostiquée via l'hétérogénéité engendrée au sein d'un groupe de gènes.

Nous proposons ici une méthode d'analyse longitudinale par groupe de gènes *Time-course Gene Set Analysis* (TcGSA) basée sur la vérification d'hypothèses utilisant des effets aléatoires pour tester directement la significativité de groupes de gènes définis *a priori*. Elle tient compte de la possible hétérogénéité des groupes de gènes et est robuste

aux designs déséquilibrés dus à des valeurs manquantes aléatoirement grâce aux estimations du maximum de vraisemblance.

Méthodes

Modèles mixtes pour l'expression d'un groupe de gènes

Soit S un groupe de gènes d'intérêt. Dans le cas d'un seul groupe de traitement (où chaque patient est son propre contrôle), l'expression génique est modélisée au cours du temps via la fonction f :

$$\text{Pour tous les gènes } g \in S, \quad y_{gpi} = \mu + \beta_g + c_{gp} + f_g(t_i) + \varepsilon_{gpi} \tag{1}$$

où y_{gpi} est l'expression du gène g du patient p mesuré au temps i, μ est le niveau moyen d'expression dans le groupe de gènes S, β_g est l'effet fixe du gène g, $c_{gp} \sim \mathcal{N}(0, \sigma_c)$ est un effet aléatoire, t_i est le $i^{\text{ème}}$ temps de mesure, $\varepsilon_{gpi} \sim \mathcal{N}(0, \sigma)$ est un terme d'erreur, et $f_g(t_i)$ est une fonction du temps (linéaire, polynomiale, splines, etc). Chaque coefficient de f est en réalité composé de deux parties : un effet fixe du temps et un effet aléatoire (permettant de tenir compte de l'hétérogénéité) du temps. Des variations autour de ce modèle sont proposées dans l'implémentation de TcGSA.

Test de la significativité d'un groupe de gènes

On veut alors tester l'ensemble des coefficients de $f_g(\cdot)$ simultanément afin de pouvoir détecter à la fois les groupes de gènes changeant de façon homogène et les groupes de gènes changeant de façon hétérogène. Un Test du Rapport de Vraisemblance (TRV) est la manière la plus naturelle de le faire, en estimant le modèle (1) sous l'hypothèse nulle $(f_g(\cdot) = 0)$ et sous l'alternative $(f_g(\cdot) \neq 0)$. Néanmoins, la distribution du TRV sous l'hypothèse nulle n'est pas triviale lorsque l'on teste simultanément plusieurs effets aléatoires et fixes. Il est possible de l'approximer [Self and Liang, 1987; Molenberghs and Verbeke, 2007] par un mélange de lois du χ^2 dépendant du nombre d'effets fixes et du nombre d'effets aléatoires à tester :

$$LRT_{H_0} \sim \sum_{k=q}^{q+r} \binom{r}{k-q} 2^{-r} \chi^2_{(k)}$$

où q est le nombre d'effets fixes et r le nombre d'effets aléatoires à tester simultanément. Habituellement, on va s'intéresser à plusieurs groupes de gènes dans une même analyse, et de nombreux TRV vont être calculés. Il est alors indispensable de corriger pour la multiplicité des tests.

Visualisation

Une fois qu'un groupe de gènes S a été identifié comme significatif, on veut résumer sa dynamique temporelle. Néanmoins, à cause de la possible hétérogénéité de S, ce n'est pas évident. Nous proposons d'identifier automatiquement le nombre de tendances différentes à l'intérieur de S grâce à la statistique *gap* développée par Tibshirani et al. [2001]. Par ailleurs, on s'intéresse souvent à un grand nombre de groupes de gènes (quelques centaines à quelques milliers) à la fois. Une conséquence supplémentaire de cette multiplicité (outre la nécessité de corriger le niveau de significativité du TRV) est de rendre la visualisation des résultats difficile. Nous proposons de représenter l'ensemble des différentes tendances que peuvent contenir chacun des groupes de gènes significatifs au sein d'une même représentation (carte de chaleur ou *heatmap*). Les tendances ayant une dynamique semblable sont rapprochées par une classification hiérarchique. Les dynamiques générales animant les données sont alors visibles, et des interactions entre diverses fonctions biologiques peuvent apparaître.

Applications

Simulations

Une étude de simulation a démontré d'excellentes propriétés de l'approche TcGSA, tant au niveau du contrôle de l'erreur de type-I que de la puissance statistique dans le cas de données longitudinales.

L'essai DALIA-1

L'essai DALIA-1 est un essai de vaccin thérapeutique contre le VIH de phase 1 (détails sur http ://clinicaltrials.gov/ct2/show/NCT00796770). Le candidat vaccin est basé sur des cellules dendritiques productrices d'interféron-α générées ex-vivo, chargées avec des lipopéptides de VIH-1, et activées avec du lipopolysaccharide. Nous nous intéressons ici aux mesures de l'expression génique durant cet essai (mesuré en sang total avec des puces *Illumina HumanHT-12 v4 Expression BeadChips*). Les patients, infectés par le VIH, ont reçu le vaccin alors qu'ils étaient traités par un traitement antirétroviral (ARV) durant la phase de vaccination, avant d'interrompre leur traitement ARV durant la seconde phase de l'essai.

Après pré-traitement des données (modèle de mélange normal-exponentiel [Shi et al., 2010], suivi de la méthode *ComBat* [Johnson et al., 2007] pour corriger les effets techniques), on dispose finalement dans cet essai de 14 temps de mesure chez 18 patients infectés par le VIH. Les données se repartissent sur deux phases distinctes, et séparées

pour cette analyse : i) la phase de vaccination (5 mesures) ; ii) la phase après l'arrêt de traitement antirétroviral (9 mesures). Dans cette analyse, nous nous sommes intéressés à des groupes de gènes relatifs au système immunitaire définis par Chaussabel et al. [2008].

La méthode TcGSA, appliquée à ces 260 modules fonctionnels, a nettement amélioré les résultats par rapport à une analyse gène-par-gène de l'expression différentielle au cours de la phase de vaccination. En effet, une analyse gène-par-gène n'avait révélé aucun changement significatif après correction pour la multiplicité des tests, alors que TcGSA a identifié 69 groupes de gènes significatifs. Par ailleurs, après l'interruption du traitement ARV, TcGSA détecte un changement significatif de nombreux modules. Ce second résultat était attendu puisque l'arrêt du traitement ARV représente une source de perturbation importante pour le système immunitaire.

Étude vaccinale

TcGSA a également été appliqué à une étude comparant le vaccin anti-pneumocoque et le vaccin anti-grippe à un placebo [Obermoser et al., 2013]. Une analyse par groupe de gènes de l'expression génique répétée suite à la vaccination, testant également les modules de Chaussabel et al. [2008], y a originellement été menée à l'aide d'un modèle linéaire plus simple (de manière transversale, sans tenir compte des données répétées). Les résultats obtenus par TcGSA d'une part ont confirmé ceux de Obermoser et al. [2013], et d'autre part les ont complété, notamment grâce à la puissance statistique accrue de l'approche TcGSA. L'un des principaux apports de TcGSA a été la double identification d'un ensemble de cinq modules (M3.2, M4.2, M4.13, M5.1 et M5.7) liés aux voies biologiques (*pathway*) de l'inflammation. Ces derniers, qui n'étaient identifiés que dans le vaccin anti-pneumocoque par Obermoser et al. [2013], ont été identifiés dans les deux vaccins par TcGSA.

Conclusion

Les mesures répétées d'expression génique sont de plus en plus courantes. Appliquée dans des études transversales, l'analyse par groupe de gènes a démontré ses qualités en termes de sensibilité et d'interprétation. Nous étendons ici cet outil aux données longitudinales d'expression génique, en tenant compte de la possible hétérogénéité des groupes de gènes. Notre approche TcGSA teste si les gènes appartenant à un groupe donné ont une expression stable au cours du temps, grâce aux estimations du maximum de vraisemblance. Cette approche peut s'appliquer dans le cas de données déséquilibrées dues à des données manquantes aléatoirement. Une classification non supervisée des dynamiques estimées pour les gènes appartenant à un groupe de gènes est ensuite réalisée, afin d'y exhiber les principales tendances. Nous avons appliqué TcGSA dans

l'essai DALIA-1, un essai de vaccin thérapeutique contre le VIH au cours duquel des patients infectés par le VIH-1 ont reçu un vaccin à base de cellules dendritiques, avant d'arrêter temporairement leur traitement antirétroviral. La méthode TcGSA, appliquée à 260 modules fonctionnels, a nettement amélioré les résultats par rapport à une analyse gène-par-gène au cours de la phase de vaccination en identifiant 69 modules aux dynamiques significatives. D'autres résultats encourageant ont également été obtenus par TcGSA sur une étude comparant deux vaccins préventifs à un placebo. Une étude de simulations a démontré la puissance statistique de l'approche TcGSA pour des données d'expression génique longitudinales. La méthode TcGSA a été implémentée sous ® dans le package TcGSA, dont la dernière version est disponible sur le CRAN repository (http://cran.r-project.org/web/packages/TcGSA/index.html).

3 Analyse intégrative de l'expression génique dans deux essais vaccinaux

Ce chapitre présente deux analyses intégratives dans le cadre d'essais vaccinaux. La première concerne l'essai de vaccin thérapeutique contre le VIH DALIA-1 déjà présenté dans la section précédente, la seconde une étude de la différence de la réponse au vaccin anti-grippale trivalent. L'enjeu de ces deux analyses est à chaque fois de révéler une partie des mécanismes sous-jacents qui expliqueraient pourquoi certains patients ont une meilleure réponse immunitaire que d'autres à la suite de la vaccination.

Analyse intégrative de l'essai DALIA-1

Au cours de l'essai DALIA-1, en plus de la mesure répétée de l'expression génique, différents marqueurs immunologiques ont été mesurés. Notamment, le maximum de la charge virale lors du rebond observé suite à l'interruption thérapeutique. Selon Lévy et al. [2014], les 16 patients suivis peuvent être séparés en deux groupes de même taille selon leur maximum de charge virale, avec d'une part ceux dont le logarithme (en base décimale) de ce maximum est inférieur à 5 et qui sont considérés comme de bons répondeurs, et d'autre part ceux pour qui il est supérieur à 5 et qui sont considérés comme de mauvais répondeurs. Afin d'identifier une signature d'expression génique associée à ce maximum d'intensité du rebond viral suite à l'interruption thérapeutique, nous avons effectué une analyse sparse Partial Least Squares (sPLS) [Le Cao et al., 2008] associant des marqueurs immunologiques à l'expression génique des 5 399 gènes participant aux modules identifiés comme actifs durant la phase de vaccination par TcGSA.

La méthode Partial Least Squares (PLS) recherche pour chacune des deux matrices de

données mises en relation, des variables latentes qui sont des combinaisons linéaires des variables originales, et qui maximisent la covariance entre elles. Cette recherche est répétée dans un processus itératif (similaire à celui de la construction des composantes principales lors d'une ACP par exemple), afin de construire des variables latentes orthogonales deux à deux pour une même matrice de données, qui maximisent à chaque fois la covariance avec la variable latente correspondante pour l'autre matrice de données. La méthode sPLS est une extension de la PLS où les variables latentes sont creuses (*sparse*), c'est-à-dire que peu de variables originales contribuent à une variable latente, grâce à l'ajout d'une pénalité de type LASSO par exemple dans la maximisation de la covariance.

On a effectué une analyse sPLS mettant en relation l'expression génique ayant évoluée au cours de la vaccination (69 modules identifiés par TcGSA) avec la production de cytokines (interféron-γ, IL-2, IL-13, I-L21), la polyfonctionnalité des cellules CD4, deux scores calculés à partir des données Luminex [Lévy et al., 2014], et le rebond viral maximum dichotomisé. Cette analyse a permis d'associer une plus faible expression des voies biologiques inflammatoires suite à la vaccination à une meilleure réponse immunitaire lors de l'interruption thérapeutique. Des analyses de sensibilité ont confirmé la signature identifiée au niveau des Modules de Chaussabel et al. [2008]

Analyse systémique des différences entre les sexes dans la réponse au vaccin anti-grippal

Le sexe est connu pour être une source importante de variabilité immunologique entre les individus. Ainsi, les hommes sont plus fréquemment sujets à des infections que les femmes, ces dernières montrant également une meilleure réponse aux vaccins [Klein, 2000; Klein and Poland, 2013]. Néanmoins les mécanismes biologiques à l'œuvre derrière ces différences sont encore mal compris. A l'heure actuelle, aucune association claire n'a été établie entre de telles différences biologiques et cliniques entre les sexes. Dans une étude, Klein et al. [2010] a montré que la plupart des gènes différentiellement exprimés lors de la vaccination contre la fièvre jaune [Gaucher et al., 2008], étaient en réalité activés préférentiellement chez les femmes. Néanmoins, ces différences n'ont pour l'instant pas été associées à de faibles taux d'anticorps.

Afin d'étudier les différences dans le système immunitaire entre les hommes et les femmes, nous avons analysé des données provenant d'une étude récente portant sur 91 individus (37 hommes et 54 femmes) répartis dans deux groupes d'âge (de 20 à 30 ans et de 60 à 89 ans) [Furman et al., 2013]. Un grand nombre de marqueurs immunologiques a été mesuré avant la vaccination dans le sang circulant tels que la production de cytokines, chemokines, diverses fréquences de populations cellulaires, ainsi que l'expression génique

en sang total. L'expression des gènes avait auparavant été résumée en 109 variables représentant 109 groupes de gènes définis à partir de ces mêmes données [Furman et al., 2013]. Par ailleurs, quatre individus furent retirés de l'analyse à cause de prélevements manquants.

La réponse au vaccin est évaluée par le rapport entre le taux d'anticorps micro-neutralisants avant et après la vaccination. Les individus sont considérés bons répondeurs si ce rapport est supérieur ou égal à quatre. Parmi les trois souches de virus contenues dans le vaccin (H1N1, H3N2, B), la différence la plus marquée entre les sexes a été observée pour la souche H3N2, sur laquelle la suite des analyses s'est concentrée.

Afin d'évaluer l'association entre le sexe et la probabilité d'être un bon ou un mauvais répondeur au vaccin pour la souche H3N2, on a utilisé un modèle de régression logistique. Dans un premier temps, deux variables potentiellement confondantes pour l'effet du sexe sur la probabilité d'être un bon répondeur ont été identifiées, à l'aide d'une stratégie ascendante d'inclusion des variables modifiant le coefficient lié au sexe de plus de 20% dans la régression logistique. Il s'agit d'un groupe de gènes lié à la production de protéines ribosomales, ainsi que du marqueur d'inflammation aiguë CRP (C Reactive Protein). Ensuite, on a appliqué une méthode innovante développée par Simon and Tibshirani [2012] pour l'identification d'interactions significatives dans un modèle linéaire où la variable réponse est binaire, qui permet d'assurer un contrôle du taux de faux positifs (*FDR*) correct dans un contexte de grande dimension. On a ainsi identifié l'effet différentiel selon le sexe d'un groupe de gènes lié au métabolisme lipidique sur la réponse au vaccin.

Finalement, il s'avère que cette interaction est d'autant plus importante lorsqu'on stratifie les hommes selon leur niveau de testostérone (une hormone dont la concentration est très basse chez les femmes) : plus leur niveau de testostérone est élevé, plus l'influence de l'expression de ce groupe de gènes sur la probabilité d'être un bon répondeur est contrastée par rapport à celle observée chez les femmes.

4 Modèles de mélange de distributions t asymétriques à processus de Dirichlet pour la modélisation de données de cytométrie en flux

La cytométrie en flux est une technologie à haut débit utilisée pour quantifier simultanément différents marqueurs cellulaires de surface et intracellulaires, à l'échelle individuelle de chaque cellule. Les développements et améliorations de cette technologie permettent aujourd'hui de décrire des millions de cellules individuellement à partir d'un échantillon sanguin, et ce sur plusieurs marqueurs. En conséquence, on obtient des jeux de données

de taille plus d'un million de fois supérieure à celle d'il y a quelques années, dont le traitement manuel se révèle très fastidieux et peu reproductible.

De nombreuses méthodes ont été développées afin de distinguer automatiquement les différentes populations cellulaires à partir de telles données [Aghaeepour et al., 2013]. Cependant, la plupart d'entre elles s'intéressent au cas d'un seul échantillon (chez un seul patient) utilisant très peu de couleurs (c'est-à-dire de marqueurs). D'autant que dans le cadre d'essais cliniques, on dispose maintenant habituellement des mesures pour une douzaine de marqueurs, mesures souvent répétées pour chaque patient à chaque temps de mesure.

Nous proposons une approche bayésienne non paramétrique pour modéliser ce type de données par des modèles de mélange à processus de Dirichlet [Ferguson, 1973; Antoniak, 1974; Lo, 1984; Escobar and West, 1995; Teh, 2010]. De tels modèles permettent d'estimer le nombre de populations cellulaires différentes sans avoir recours à des outils de sélection de modèle destinés à choisir le modèle comprenant le nombre de populations cellulaires le plus adapté aux données. Notre modélisation étend le modèle de mélange gaussien à processus de Dirichlet à la distribution skew t [Azzalini and Capitanio, 2003] (asymétrique et à queue lourde), basée sur la paramétrisation de Frühwirth-Schnatter and Pyne [2010]. Si C est le nombre de cellules observées, et y_c désigne l'observation des différents marqueurs cellulaires pour la cellule c, et k est l'indice de la population cellulaire, notre modèle peut s'écrire :

$$\alpha | a, b \sim \text{Gamma}(a, b)$$
$$\boldsymbol{\pi} \mid \alpha \sim \text{GEM}(\alpha)$$

pour $k = 1, 2, \ldots$

$$\xi_k, \psi_k, \Sigma_k, \nu_k \sim G_0$$

pour $c = 1, 2, \ldots, C$

$$\ell_c \mid \boldsymbol{\pi} \sim \text{Mult}(\boldsymbol{\pi})$$
$$\gamma_c \mid \ell_c, (\nu_k) \sim \text{Gamma}\left(\frac{\nu_{\ell_c}}{2}, \frac{\nu_{\ell_c}}{2}\right)$$
$$s_c \mid \gamma_c \sim \mathcal{N}_{[0,+\infty[}\left(0, \frac{1}{\gamma_c}\right)$$
$$y_c \mid \ell_c, \gamma_c, s_c, (\xi_k, \psi_k, \Sigma_k) \sim \mathcal{N}\left(\xi_{\ell_c} + \psi_{\ell_c} s_c, \frac{1}{\gamma_c}\Sigma_{\ell_c}\right)$$

Un tel modèle est particulièrement approprié pour la classification non supervisée de données de cytométrie en flux grâce aux trois caractéristiques suivantes : i) il évite de fixer

le nombre de populations à identifier en l'estimant directement à partir des données; ii) les données de cytométrie en flux sont particulièrement bien représentées par des distributions asymétriques et à queue lourde telles que la skew t (dont la loi normale est un cas particulier); iii) il produit une estimation de la probabilité a posteriori de co-classification pour chaque paire d'observations, permettant ainsi une quantification de l'incertitude autour de la partition des données. Nous avons développé un algorithme d'échantillonnage de Gibbs partiellement replié (*partially collapsed* [van Dyk and Park, 2008; van Dyk and Jiao, 2014]) basé sur l'échantillonnage par tranche (*slice sampling* [Neal, 2003; Walker, 2007; Kalli et al., 2011] pour estimer de tels modèles. Nous l'avons implémenté en⬡ sous la forme d'un package, NPflow, qui est actuellement encore en développement.

Dans le cas où les mesures de cytométrie en flux sont répétées au cours de l'étude, nous proposons une stratégie d'approximations séquentielles du posterior (sous l'hypothèse que l'ensemble des données proviennent de la même distribution), en traitant les échantillons répétés les uns à la suite des autres, lors de l'acquisition des données. Cela permet d'utiliser l'information a priori obtenue sur les échantillons précédents pour pouvoir traiter l'échantillon courant, sans avoir à attendre la fin de l'étude et de devoir traiter l'ensemble des données simultanément, à l'inverse d'approches hiérarchiques telles que celles proposées par Cron et al. [2013] ou Dundar et al. [2014].

En comparant nos résultats sur des jeux de données réelles utilisés dans une compétition visant à comparer un grand nombre d'approches de traitement automatique de données de cytométrie en flux, nous avons obtenu des résultats dans la moyenne des différents algorithmes proposés. Sur les simulations, la stratégie d'estimation séquentielle du posterior offre des résultats prometteurs. Néanmoins, il faut reconnaître que l'évaluation rigoureuse des algorithmes de traitement automatiques des données de cytométrie en flux est rendue difficile en l'absence d'une véritable référence (*gold-standard*) pour les données réelles. En effet, à l'heure actuelle on utilise le traitement manuel comme référence malgré le fait avéré qu'il soit relativement variable d'un opérateur à l'autre, peu reproductible et bien souvent ne couvre pas tout l'espace des observations [Ge and Sealfon, 2012; Aghaeepour et al., 2013]. Une solution pourrait être d'utiliser un consensus entre différents traitements, manuels ou automatiques, des données à la place d'un seul traitement manuel [Aghaeepour et al., 2013].

5 Discussion générale

Grâce à de nombreuses améliorations technologiques, les données aujourd'hui générées dans la recherche vaccinale permettent d'étudier de manière très précise les cellules prélevées, de l'expression de leurs gènes à leur production de cytokines, en passant par leurs

26

marqueurs de surface. Ces données sont intrinsèquement de grande dimension, même au niveau des populations cellulaires. Au cours de ce travail, nous nous sommes efforcés de proposer de nouveaux outils statistiques afin d'améliorer l'analyse de telles données et d'intégrer différents niveaux biologiques au sein d'une même analyse. Une idée prépondérante qui sous-tend l'ensemble de ce travail est la recherche de l'utilisation au maximum de l'ensemble de l'information disponible. Elle se fait d'une part en incorporant de l'information a priori dans la modélisation statistique, et d'autre part en intégrant toutes les données mesurées, dans l'espoir d'améliorer les résultats d'inférence et leur interprétation. Cette ambition est motivée par le renforcement des mesures longitudinales dans la recherche biologique et clinique, y compris pour les données de grande dimension. Ainsi, à la fois pour les données d'expression génique, et pour les données de cytométrie en flux, nous avons développé de nouvelles approches statistiques afin d'essayer de tirer partie de telles mesures répétées.

Nous avons proposé une approche d'analyse par groupe de gènes spécialement centrée sur les mesures longitudinales, qui permet de répondre avec une puissance statistique optimale à la question « Quels sont les groupes de gènes dont la dynamique d'expression évolue au cours de l'étude ? », comblant certaines lacunes par rapport aux approches similaires de la littérature [Subramanian et al., 2005; Efron and Tibshirani, 2007; Hummel et al., 2008; Shahbaba et al., 2011; Wu and Smyth, 2012]. Néanmoins, dans le cas d'analyses par groupes de gènes, une attention toute particulière doit être portée aux groupes de gènes testés, car ceux-ci constituent en réalité le premier niveau d'hypothèse du modèle. Si dans ce travail, nous les avons considérés connus a priori, il existe en réalité un veritable domaine de recherche pour l'inférence de voies biologiques à partir de données temporelles d'expression génique [Wu et al., 2014; Ratmann et al., 2009]. Par ailleurs, nous nous sommes focalisés sur les données de puce à ADN dans ce travail, mais il existe d'autres façons de mesurer l'expression génique. Par exemple les méthodes de séquençage direct de l'ARN, bien plus précises que les puces à ADN. Si leur traitement fait encore l'objet de développements actifs à la recherche de méthodes appropriées, il semble néanmoins que notre méthode TcGSA soit facilement transposable à ce type de données en se basant sur l'approche de Law et al. [2014]. Il s'agit simplement de modéliser directement le logarithme de la proportion d'ARN mesuré tout en tenant compte de l'hétéroscédasticité à l'aide de pondérations de la précision.

Concernant notre modélisation des données de cytométrie en flux par des modèles de mélange de distributions skew t à processus de Dirichlet, plutôt que de nous tourner vers une approche hiérarchique telle que celles proposées par Cron et al. [2013] ou [Dundar et al., 2014], nous avons choisi une stratégie d'estimation séquentielle qui ne requiert pas de traiter l'ensemble des données en une seule estimation. Face à l'augmentation des tailles

d'échantillons, notamment due à des mesures longitudinales, nous pensons que c'est un avantage en faveur de notre approche. Les perspectives de développement et d'application de ce travail sont importantes notamment par l'utilisation à la fois sur des données de moins grande dimension recueillie au cours du suivi des patients hospitalisés et sur des données de plus grandes dimensions avec les nouvelles techniques de spectrométrie de masse (CytoF).

Enfin, si nous avons développé deux exemples d'analyses intégratives dans cette thèse, qui ont chacune conduit à formuler de nouvelles hypothèses sur les processus biologiques sous-jacents, une étape supplémentaire serait d'utiliser les données de cytométrie en flux afin de déconvoluer l'expression génique (plus correctement l'abondance génique) avec les variations de population cellulaire, à la manière de Shen-Orr et al. [2010] mais au cours du temps. Cela permettrait de faire la distinction entre une variation d'abondance due à la circulation de certaines populations cellulaires spécifiques, et une variation d'abondance due à un réel changement d'expression au niveau des cellules.

L'analyse du déluge de données qu'est en train de connaître la recherche biologique et clinique nécessite de nouvelles méthodes statistiques, qui sont en train d'être développées. De telles méthodes sont souvent spécifiquement adaptées à chaque question d'analyse, et donc aux données disponibles. Ces données, qui sont les observations complexes d'un système encore plus complexe, nécessitent des modélisations sophistiquées, interdisant le plus souvent l'utilisation de statistiques éprouvées qui se révèlent trop simplistes et ne permettent pas de répondre à la question posée. Ce phénomène offre aux biostatisticiens des perspectives de recherche ambitieuses, à une époque où les collaborations translationnelles sont devenues indispensables.

1 Introduction

1.1 The analysis of high dimensional data in clinical research

Statistics for high-dimensional data has been a field of growing interest for a few decades, including biostatistics for high-dimensional data (Figure 1.1). More recently, the term "Big Data" has been coined to designate the flow of high dimensional data generated in various contexts. We have definitely entered the "Big Data" era, as one can see from a search on Google Trends (Figure 1.2). Apart from the current hype that surrounds this key words, the real underlying phenomenon at play is the massive production of data, at a rate that keeps increasing. The main data providers are the usuals suspects: particle physics (data from particle colliders), astronomy (data from high definition telescopes such as Planck), Facebook, YouTube, emails... But medicine and biology are also involved in the increasing amount of data generated [Marx, 2013]. At the end of the twentieth century, the first human genome took more than 10 years to be sequenced, with an associated cost exceeding 3 billion dollars. Today, 10 years later, a whole genome can be sequenced within the day at a cost not exceeding 1,000 dollars. Technological headways have made data much cheaper and much easier to get. In addition to sequencing and other genomics data, every biomedical data type is concerned: from peptides (proteomics), to imaging (fMRI data), to cellular markers (high-throughput flow-cytometry). All these biomedical big – or high-dimensional – data have lead to the publication of an always increasing number of related articles indexed on Pubmed (Figure 1.1). Moreover, because of the need for tackling the issues that arise from such data, the development of methods for high-dimensional data is soaring (even though "high-dimensional" data is not brand new in biostatistics, the first occurrences of this term in the literature dating back to the mid-seventies). But the challenge does not only concern the volume of data. The increase in the amount of information is coupled with an increase in complexity of the data, generated from different sources. The resulting requirement of integrating such a high variety of data constitutes a key challenge in the analysis of big data.

The increase of collected data also affects clinical trials [Thiébaut et al., 2014]. The measurement of whole genome expression has become more and more common in human
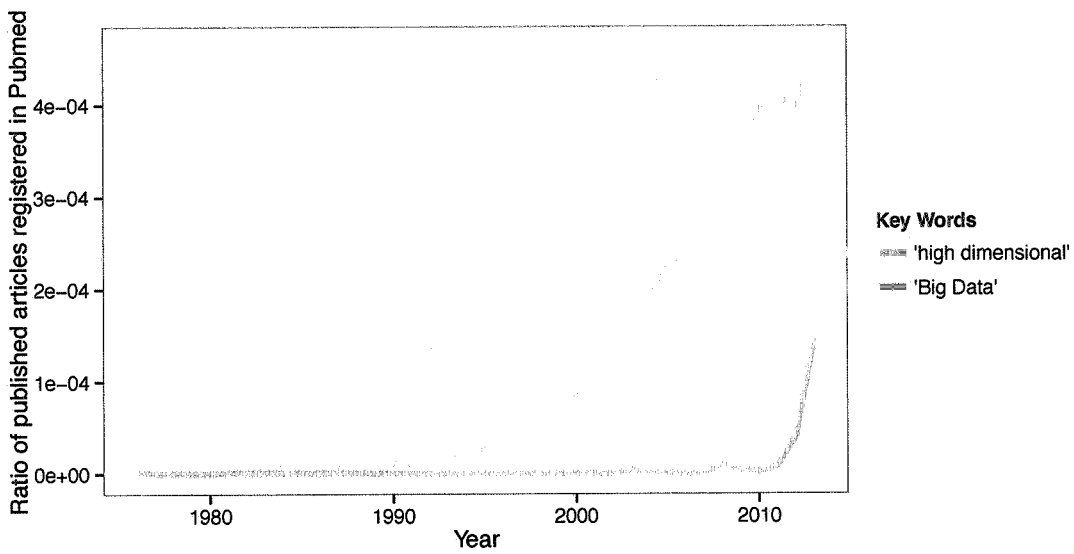
Figure 1.1 – Trends for the key words 'Big Data' and 'high dimensional' on PubMed
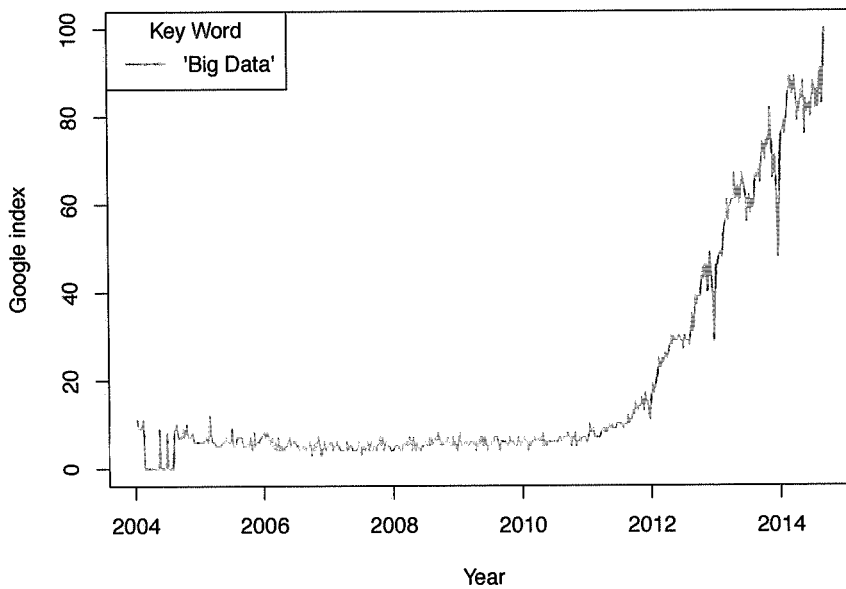
Figure 1.2 – Trend for the key word 'Big Data' on Google Trends

clinical trials, sometimes at repeated time points. In addition to genomics measures, imaging, flow cytometry and other high-throughput technology measures can often be part of clinical measurements, leading to the idea of deep phenotyping of individuals [Tracy, 2008; Robinson, 2012]. All these add up to create real big clinical trial data. The particularity of this context is that the high-dimension of the data concerns each patients. This is somewhat different from the context of the big data observed in the web or in other industries for instance. The question however remains how to deal with such amounts of data. In exploratory analyses, the number of potential associations to be investigated rapidly increases as with data dimensionality, raising multiplicity issues. If analysis is not done carefully, the number of false positive grows very rapidly. Moreover, those high dimensional data are often generated from different sources, different tissues, and measure various aspects of a biological reality. Indeed, gene expression, intracellular cytokine production, cell surface markers, etc. can be measured for the same patient. Analyzing all those different kinds of data simultaneously for the same individual is very complex. The data generated through this deep phenotyping of individuals needs to be integrated to answer a given research question. On the one hand, the information carried by all the data altogether is often largely superior to the noise of the numerous measurements. On the other hand, a thorough analysis of all these data requires sophisticated statistical tools and thoughtful analysis strategy. Furthermore, the high dimensionality implies substantial computation times.

Most of the works presented in this thesis emanate from a specific therapeutic HIV vaccine trial: the DALIA-1 trial. It is a therapeutic HIV vaccine clinical trial in which 19 patients were vaccinated. This trial evaluated the administration of a dendritic cell based vaccine to HIV infected patients as a way to boost their immune response against HIV infection. In order to better understand the underlying biological mechanisms activated by this vaccine, a huge number of data were collected during this trial: longitudinal gene expression in the blood was repeatedly measured with microarrays over the course of the trial, as well as blood cell markers that were measured with flow cytometry and multiplex technologies.

1.2   The search for an effective HIV vaccine

Since the beginning of the pandemic, HIV/AIDS has caused more than 35 (95% CI: 35-43) million deaths. Today, 35.3 (95% CI: 32.2-38.8) million people are living with HIV and 2.3 (1.9-2.7) million new cases of HIV infection occur per year [The Joint United Nations Programme on HIV/AIDS (UNAIDS), 2013, 2014]. The development of highly active antiretroviral therapies has improved the prognostic of infected individuals [Lewden

et al., 2007]. It may also impact the pandemic by reducing the transmission of the virus [Granich et al., 2009]. Several approaches have already demonstrated their efficacy such as male circumcision [Auvert et al., 2005] or antiretroviral prophylaxis [Abdool Karim et al., 2010; Grant et al., 2010]. Vaccination is usually the most effective intervention to prevent and control infectious diseases but the development of HIV vaccine remains challenging in particular because of the high variability of its genomic sequence [Rappuoli and Aderem, 2011].

Promising advances in HIV vaccine development include novel approaches in immunization strategies, including prime-boost immunization with heterologous vectors (e.g. attenuated viral vectors, protein-based vaccines) and new methods for antigen presentation. The first positive results from a phase 3 trial (the RV144 Trial) showing a vaccine efficacy of 31% has underlined the relevance of prime-boost strategies [Rerks-Ngarm et al., 2009]. Researchers are now developing and comparing several prime boost strategies using different combinations of candidates to ultimately increase the vaccine efficacy. New ways of antigen presentation are also emerging. Immunologists are now able to load HIV-1 antigens in the best antigen presenting cells, i.e. dendritic cells [Cobb et al., 2011]. Such a monocyte-derived dendritic cell vaccine is currently evaluated in phase 1 trials such as the DALIA-1 trial. One approach to improve delivery of protein vaccines to dendritic cells is to introduce the protein into monoclonal antibodies that efficiently target dendritic cell receptors. In macaques, the immunological response has been then showed to improve [Flynn et al., 2011], and it could be oriented according to the targeted dendritic cell receptor[Cobb et al., 2011; Flamar et al., 2013].

Vaccines are a corner stone of modern medicine. It is by no comparison the single most effective intervention against an infectious disease [Pulendran and Ahmed, 2011]. One of the main challenge in developing an effective HIV vaccine is the identification of predictive correlates of immunity [Roederer et al., 2014]. Indeed, if the first vaccines date back to more than 200 years, the details of the immunological pathways triggered by vaccines still remains partly unknown [Pulendran and Ahmed, 2011]. Hopefully, the late downpour of data faced in clinical trials will help to improve this knowledge. Nevertheless analyzing this stream of complex data requires new approaches suited for this high-dimensional, versatile context. Recently, systems biology approaches have been developed in order to integrate global sets of biological data from many hierarchical levels. Their goal is to identify emergent properties that are not demonstrated and cannot be predicted from their individual parts alone [Zak and Aderem, 2009]. The system analysis of vaccine usually aims at finding signatures that are predictive of protection, but they can also provide insights into the mechanisms underlying protection. A good example is the yellow fever vaccine. A recent microarray analysis gave signatures taken 3 and 7 days after yellow

fever vaccination that are able to predict B- and T-cell responses measured at a later time [Querec et al., 2008]. Gene expression analysis also highlighted the activation of some components of the immune system that were not expected, and are not usually explored phenotypically, such as the innate immune system.

1.3   Methodological challenges

In high-dimensional settings, standard modeling tools, such as multivariate linear models for instance, are usually not identifiable. Solutions to methodological challenges faced with high-dimensional data include resorting to (over-) simplistic univariate modeling strategies, correcting for the multiplicity downstream of the analysis, performing variable selection, or adding prior biological knowledge to reduce the model complexity. In multivariate linear models, high-dimensionality generally prevents the use of Ordinary Least Squares (OLS) estimators (more details are presented in Appendix C page 131). Indeed co-linearities between covariates quickly arises as their number increase. In the extreme but common case where the number of individuals is lower than the number of covariates considered, the OLS. One can still perform a regression analysis by penalizing the likelihood to optimize. Different penalties are available, relying on different norms of the regression coefficient vector: for instance the LASSO penalty (L^1 norm) [Tibshirani, 1996], the ridge penalty (L^2 norm) [Hoerl and Kennard, 1970] or the elastic net (balance between L^1 and L^2 norms) [Zou and Hastie, 2005].

One might therefore consider high dimensionality only as an obstacle to overcome in an analysis process. However, high dimensionality can revealed itself as a major asset in separating noise from signal. As an example, VSURF [Genuer et al., 2010] leverages the high dimensionality of the data to estimate the level of variability of a noise variable. Besides, many multiple testing correction methods require to estimate the proportion of true positive as precisely as possible [Guedj et al., 2009]. In both cases the high-dimensionality of the data becomes a strength for the estimation procedure. High dimensionality of the data can in fact be both a complication and a helper regarding the analysis.

High-dimensional data are often structured data. For example the fact that gene expression likely precedes cellular cytokines production induces a structure between the gene expression variables and cytokines production variables. In addition an important amount of prior knowledge can be available; this prior knowledge can sometimes help to identify even deeper structures, such as in the case of gene expression data with prior information taking the form of pathway databases like Gene Ontology [Ashburner et al., 2000] or KEGG [Kanehisa and Goto, 2000]. Taking into account the structure and the prior knowledge can lead to significant improvement in the modeling strategy, for example

by influencing a variable selection procedure, or by decreasing the loss of statistical power due to multiplicity correction, etc.

Another recent challenge appeared through the repetition of the measurements by high throughput assays. This constitutes a great opportunity to better disentangle the relationship between markers, but requires specific approaches that take into account the hierarchical structure of the data. However, being able to separate the within and between individual variability might be very informative. For instance, it can help in identifying groups of patients for which the response to the intervention differs, leading to the potential of personalized medicine. As a practical example, sex can be an important factor influencing the immune response to vaccine [Klein et al., 2010; Klein and Pekosz, 2014].

An integrative analysis of high dimensional data is all the more powerful than prior knowledge and natural structure of the data are used, while as much heterogeneity as possible is taken into account. However, depending on which modeling strategy is chosen, some of these goals can be relatively difficult to achieve.

1.4   Thesis objective and outline

The work described in this document is primarily motivated by this new surge of data, especially in the context of clinical trials. Data that cannot be ignored, but whose high-dimensionality requires more complex analyses. The level of gene expression is well recognized as a high dimensional level that needs specific statistical tools for its sound analysis. However, phenotypic measures in vaccine trials are also high dimensional. Immune cell populations are measured by 8 to 16-color cytometers making it theoretically possibile to distinguish up to 2^{16} types of cells. Moreover such measurements are often repeated over time, during the follow-up of patients. Hence, an integrative analysis should relate these noisy high dimensional data, that is transcriptomics with phenotypic data.

First, this thesis presents an original methodological development for the analysis of longitudinal gene expression data, taking into account prior biological knowledge in the form of predefined gene sets. Then, the thesis focuses on two integrative analyses performed on two different vaccine trials, against HIV and against flu, respectively. Finally, the thesis introduces a new model-based clustering approach for the automated treatment of cell populations from flow-cytometry data, namely a Dirichlet process mixture of skew t-distributions.

2 Time course gene set analysis

Abstract: Gene expression measurements have revolutionized the way to monitor biological activity among living organisms. Thanks to microarray technology, whole genome gene expression can be measured from a simple blood sample.

After specific preprocessing of gene expression data, standard univariate gene-by-gene analysis often suffer from a lack of power due to multiplicity correction. In order to gain in statistical power, gene set analysis methods use prior biological knowledge to analyze such gene expression data. This prior knowledge takes the form of predefined groups of genes, linked through their biological function. Gene set analysis methods results are more sensitive and interpretable than those of methods investigating genomic data one gene at a time, and they have been successfully applied in cross-sectional studies.

The time-course gene set analysis (TcGSA) introduced here is an extension of such gene set analysis to longitudinal data. This method identifies *a priori* defined groups of genes whose expression is not stable over time, taking into account the potential heterogeneity between patients and between genes. When biological conditions are compared, it identifies the gene sets that have different expression dynamics according to these conditions. Data from two studies are analyzed: data from an HIV therapeutic vaccine trial, and data from a recent study on influenza and pneumococcal vaccines. In both cases, TcGSA provided new insights thanks to an increased sensitivity compared to standard approaches. Those results highlight the benefits of the TcGSA method for analyzing gene expression dynamics.

Key Words: Gene expression; Gene set analysis; Likelihood, ratio test; Linear mixed effects model; Longitudinal data;

Contents

Valorisation: section 2.3 is mainly part of an article that was submitted for publication in Plos Computational Biology and that is currently under revision.

2.1 Gene expression data

Genes are the vector of heredity. From one cell to another, from one organism to its descendants, they constitute the information that is passed across generations. It is encoded through Deoxyribonucleic Acid (DNA). In the human body, each cell contains the whole DNA of a person, encapsulated inside its nucleus. Yet, differentiation processes results in many different types of cell inside the body, with very different functions. In spite of having the same genome, those cells are different because they express different genes.

Gene expression is a dynamic process at the root of any metabolic reaction. The expression of the information encoded in the genes occurs in two steps: i) transcription, during which DNA is transcribed into messenger ribonucleic acid (mRNA); ii) translation, during which mRNA is translated to produce a protein (Figure 2.1). DNA is a double stranded polymer with a double helix shape, made of four basic nucleotides (adenine, cytosine, guanine and thymine, respectively denoted A, C, G and T). mRNA is a single stranded molecule. It is complementary to the nucleotid sequence of the DNA molecule (where the uracile nucleotide, denoted U, replaces the thymine). Finally, proteins are made of 20 different amino acids, each being coded by a specific triplet of 3 consecutive nucleotides from mRNA [Dudoit et al., 2002].

DNA   **mRNA**   **Protein**
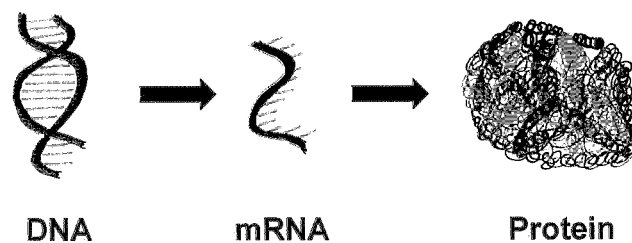
Figure 2.1 – Representation of the metabolic production chain

Gene expression can therefore be studied at many different levels. Microarrays are a convenient way to measure mRNA across the whole genome. Microarrays are made of microscopic spots attached on a solid surface (often glass or silicon). Each spot is filled with a specific nucleic acid (DNA or RNA) sequence. Those specific sequences are carefully engineered, in order to hybridize to characterized gene or candidate gene transcripts. Hybridization between two DNA (or two RNA) strands is the property of a nucleic acid sequence to specifically match only with its complementary sequence (by forming hydrogen bonds between them). Their size allows genome-wide transcriptional coverage, as tens of thousands of transcripts are profiled on a single microarray.

Numerous industrial companies market gene expression microarrays that measures

mRNA levels in a biological sample (Illumina®, Affymetrix®, Agilent...). Although microarrays may involve slightly different technologies, all rely on hybridization of labeled nucleic acid (either RNA or DNA) from the biological samples to predefined complementary probes fixed on the microarray [Schulze and Downward, 2001; Patel, 2008]. As an example, the Illumina® technology is based on the following steps: i) mRNA is extracted from the biological samples through purification; ii) this mRNA is then converted to double stranded DNA through reverse transcription; iii) this DNA is then amplified into complementary RNA (cRNA) through in vitro transcription using biotin labeled nucleotides; iv) this cRNA is then hybridized with the probes present on the array (Figure 2.2); vi) the excess of cRNA that has not hybridized with any probes is washed away; vii) biotin is bound to a fluorescent molecule (Cy3 fluorescent dye via streptavidin bound) viii) Fluorescence emission is measured with a laser scanner [Strachan and Read, 1999; Illumina, 2010].
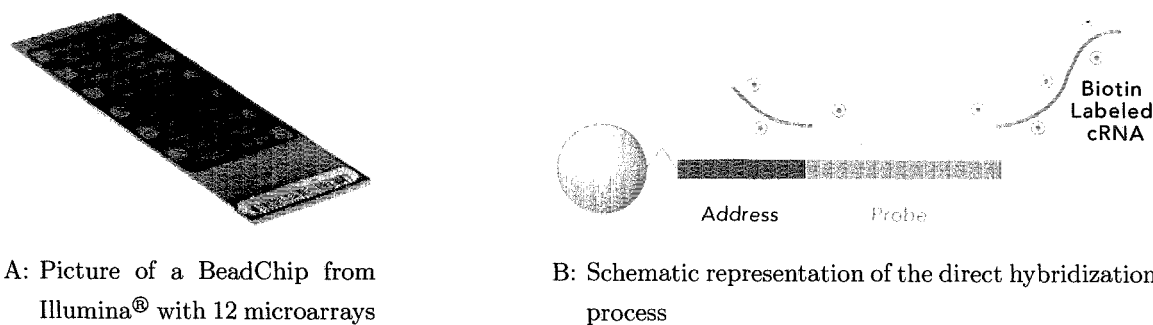
A: Picture of a BeadChip from Illumina® with 12 microarrays

B: Schematic representation of the direct hybridization process

Source: Illumina, Inc.

Figure 2.2 – Illumina® HumanHT-12 v4.0 Expression BeadChip

The gene expression could be measured on whole blood as on specific cells after selection, this latter process being much more expensive and complex. The gene expression in whole blood or peripheral blood mononuclear cells (PBMC) would be more appropriately called gene abundance, as the variation of the measures could be due to either a variation of gene expression by cells, or a variation of the number of cells expressing a given gene. Gene expression data are usually high-dimensional data. Indeed, a gene expression microarray generally measures a few tens of thousands of probes (for instance, Illumina® HumanHT-12 v4.0 Expression BeadChip arrays target a little bit more than 47,000 transcripts). As of today, a gene expression microarray costs a little less than US$100 per sample (depending on the manufacturer), so most of the time there are more probes measured than individuals ($"p >> n"$ kind of problem).

2.2 Gene expression analysis of microarray data

2.2.1 Gene expression data pre-processing

Gene expression is especially prone to technical variability [Hartemink et al., 2001]. It is therefore mandatory to preprocess such data, in order to ensure the comparability of the different samples measured.

The very first step when dealing with gene expression data is the quality control of the data. First and foremost, one generally has to make sure that enough RNA was available from each biological sample, and that its integrity was not low [Schroeder et al., 2006]. Then, various graphics tools can be useful for identifying outlier samples, whose signal intensity distribution across all the probes is significantly different compared to the other samples [Irizarry et al., 2003].

Background correction Gene expression microarrays are systematically subject to some background noise signal. There are various methods to correct the data for this background noise, but those are highly dependent on the microarray technology used [Wu et al., 2004; Shi et al., 2010]. For the Illumina plateform, numerous negative control probes (that do not bind to any known cRNA) are used to measure the background level. Then a normal-exponential convolution model can be fitted to estimate which part of the observed signal is attributable to background noise and which part constitutes the actual biological signal [Xie et al., 2009].

Normalization In order to stabilize the variance of the signal in regards of the intensity, it is desirable to reduce the scale of the data [Shi et al., 2010], for instance by applying a log2 transformation (a small offset can be added to all the data prior to this transformation). At this point there is usually still a lot of unwanted variability between the different samples of a study. Under the hypothesis that only a few probes are highly expressed, regardless of the sample, a quantile normalization [Bolstad et al., 2003] can be applied to reduce the between-sample variability. Such a normalization makes the assumption that the gene expression signal overall distribution should be the same from one sample to another. If the quantile normalization assumption is too strong or not suitable for a specific context (for instance with different tissues or in case of acute infection), loess normalization [Yang et al., 2002] (based on robust local regression) can be used instead. Both Bolstad et al. [2003] and Shi et al. [2010] perfom comparison of various normalization strategies.

Batch effect removal Finally, potential batch effects must be investigated (and corrected when possible). Microarrays are very sensitive to batch effects as their measurements are easily affected by experimental conditions. Such batch effects can have a substantial impact on the final results Leek et al. [2010]; Parker and Leek [2012].

Several methods exist for the quantification and identification of batch effect, such as the PVCA [Boedigheimer et al., 2008; Li et al., 2009] or the more recent PC-PR2 [Fages et al., 2014]. Both of above mentioned methods rely on a two step procedure: i) a principal component analysis of the expression data; ii) either a variance component analysis via multivariate linear mixed effect modeling of each component, or partial R^2 computation via a multivariate linear regression. PC-PR2 determines the variability imputable to non categorical variables or technical variables with only a few batches to be correctly estimated compared to the PVCA, and takes into account some correlation between the explanatory variables.

Once batch effects are identified, it is desirable to correct the signal for those unwanted batch effects. Several methods are available to remove batch effects [Chen et al., 2011]. Two approaches seems particularly efficient: on the one hand the ComBat algorithm, which is based on empirical Bayesian estimation of a location/scale model [Johnson et al., 2007]; and on the other hand, the Surrogate Variable Analysis (SVA), which is based on surrogate variables identification and construction from the estimated residual data once the primary effect of interest is removed [Leek and Storey, 2007]. The two methods takes two rather different approaches to the problem of batch effect: the ComBat algorithm aims at "correcting" the data for batch effects; whereas SVA constructs surrogate variables for inferred batch effects and any downstream analyzes must be adjusted on those surrogate variables to take into account batch effects. The ComBat algorithm has the advantage of taking into account potential batch effects on the variance of the signal. It can use either parametric Normal and inverse-Gamma prior distributions for the batch effects respectively on mean and variance, or non-parametric prior distributions. In both case prior distributions are estimated from the data in an empirical Bayes manner. The SVA algorithm is supposed to be able to identify batch effect even when the batch variable is unknown. Both methods face difficulties when batches are collinear with variables of biological interest.

2.2.2   Standard differential expression analysis

The most common approach to deal with the high dimensionality of gene expression data (e.g. 47,000 transcripts by sample) is to perform a univariate analysis for each probe on pre-processed data. If the results of these analyses are summarized as p-values (for

instance an association test is performed for each probes with a biological variable of interest), then those p-values can be corrected for the multiple testing issue (Appendix A page 121).

A popular and efficient method to derive such probe wise p-values is to use empirical Bayes estimation of linear models [Smyth, 2004]. An empirical Bayes model refers to Bayesian model in which the prior parameters are estimated from the data instead of reflecting expert prior knowledge (which in practice, often results in non-informative conjugate priors) [Casella, 1985; Efron, 2014].

The results of such differential expression analyses typically take the form of lists of differentially expressed genes between different biological conditions. This can have two shortcomings: first such lists can be hard to interpret (especially if a lot of genes are significantly differentially expressed), as genes; second, only strong individual gene signals can be picked up, due to the (much needed, but conservative) multiplicity correction.

2.2.3 Gene set analysis

Gene set analyses focus on predefined sets of genes, that are linked *a priori* by their biological function or their co-expression in given biological settings. Gene set analyses are more powerful than gene-by-gene univariate analyses [Subramanian et al., 2005] as they use more information. Furthermore, it is biologically more relevant to detect the small but concomitant variation of the expression of several genes from a given pathway than an intense variation of a couple of genes. Results from gene set analyses are also easier to interpret as they are usually directly annotated with relevant biological functions.

Several methods available to perform gene set analysis, that can be separated into two groups:

i) *enrichment analyses* are two steps procedures, first requiring univariate analysis of the data, and then using each probe to determine if a given set of genes is over represented among differentially expressed genes; [Subramanian et al., 2005; Barry et al., 2005; Efron and Tibshirani, 2007]

ii) *direct gene set analyses* assess the significance of a given gene set in a single step; [Kim and Volsky, 2005; Hummel et al., 2008; Wang et al., 2009; Shahbaba et al., 2011]

The enrichment procedures, despite their "top-down" approach of the problem [Liu et al., 2007], are the most popular, notably due to the commercial software Ingenuity® Pathway Analysis, and other "point & click" software that implements enrichment approaches. Another classification of the methods can be made on the type of hypothesis that is tested to identify significant gene sets (subsection 2.3.1).

One of the key step in a gene set analysis is the gene set definition. Several gene set databases exist: on the one hand Gene Ontology [Ashburner et al., 2000] or KEGG [Kanehisa and Goto, 2000] are widely used system-wide gene sets definitions, driven by biological knowledge ; on the other hand data driven immunological gene sets have been derived by Chaussabel et al. [2008] (Modules) and Li et al. [2013] (BTM) based on co-expression across multiple biological conditions. According to Li et al. [2013] data driven gene sets such as the BTM are more sensitive than knowledge driven gene sets such as KEGG or GO. The Modules and the BTM are similar in their principle, their main difference being that Modules were derived from data concerning 9 immune related pathologies produced with the Illumina® technology, whereas BTM were obtained from public datasets produced on different platforms. Numerous other knowledge based database exist: PANTHER is focused on proteins and their related genes, based on phylogenetic trees extrapolations [Mi et al., 2013] it takes advantage of experiments across different organisms to increase biological knowledge; the Reactome Knowledgebase also focuses on protein molecular functions and reactions [Croft et al., 2014], etc. Finally gene set analyses can also be performed with data-driven gene set defined directly on the same observed expression through co-expression.

2.3 Time course gene set expression analysis

This section is mainly part of an article that was submitted for publication in Plos Computational Biology and that is currently under revision.

Gene set analysis methods, which consider predefined groups of genes in the analysis of genomic data, have been successfully applied for analyzing gene expression data in cross-sectional studies. The time-course gene set analysis (TcGSA) introduced here is an extension of gene set analysis to longitudinal data. The proposed method relies on random effects modeling with maximum likelihood estimates. It allows to use all available repeated measurements while dealing with unbalanced data due to missing at random (MAR) measurements. TcGSA is a hypothesis driven method that identifies *a priori* defined gene sets with significant expression variations over time, taking into account the potential heterogeneity of expression within gene sets. When biological conditions are compared, the method indicates if the time patterns of gene sets significantly differ according to these conditions. The interest of the method is illustrated by its application to two real life datasets: an HIV therapeutic vaccine trial (DALIA-1 trial) where microarray data were measured every 4 weeks, and data from a recent study on influenza and pneumococcal vaccines where microarray data where available at day 1, 3, 7 and 21. In the DALIA-1

trial TcGSA revealed a significant change in gene expression over time within 69 gene sets during vaccination, while a standard univariate individual gene analysis corrected for multiple testing as well as a standard a Gene Set Enrichment Analysis (GSEA) for time series both failed to detect any significant pattern change over time. When applied to the second illustrative data set, TcGSA allowed the identification of 4 gene sets initially linked only with the pneumococcal vaccine, as also highly significant with the influenza vaccine. In our simulation study TcGSA exhibits good statistical properties, and an increased power compared to other approaches for analyzing time-course expression patterns of gene sets.

2.3.1 Introduction to time-course gene set analysis

Microarray experiments are increasingly used for evaluating changes in gene expression over time. The analysis of the temporal change of gene expression should help in understanding the complex mechanisms of gene regulation. For instance, transcriptional profiles have been repeatedly measured to study the change in gene expression during the natural history of SIV/HIV infection [Bécavin et al., 2011; Bosinger et al., 2012] or to evaluate the effect of vaccines [Querec et al., 2008; Palermo et al., 2011]. In the applications considered here in section 2.3.3 page 52, the investigators wanted to detect the genes for which the abundance changed over time after a vaccination (against HIV, influenza or pneumococcus)[Lévy et al., 2014; Obermoser et al., 2013].

In order to analyze such longitudinal high-dimensional data, several approaches have been suggested including a gene-by-gene statistical analysis [Storey et al., 2005; Berk et al., 2012], dimension reduction methods [Liquet et al., 2012] or gene set analysis [Wang et al., 2009]. A gene set is a group of genes that are *a priori* co-regulated or functionally linked. Examples of such gene set relating to biological processes or pathways are those defined by KEGG [Kanehisa and Goto, 2000], Gene Ontology [Ashburner et al., 2000] or Chaussabel's functional modules [Chaussabel et al., 2008]. The gene set analysis [Subramanian et al., 2005; Efron and Tibshirani, 2007; Maciejewski, 2014] is supposed to be more powerful than a gene-by-gene analysis because it can detect a change of expression of a group of genes although none of them show a very high absolute fold change. Furthermore, a change of all genes in a given pathway may be biologically more meaningful than a large increase of a single gene. Also, provided that the gene sets are well defined, the result should be more sound and comparable across studies than a gene-by-gene analysis [Subramanian et al., 2005]. Finally, gene set analysis avoids a second step for a global interpretation as described in the "bottom up" approach [Liu et al., 2007; Wang et al., 2009].

The analysis of longitudinal microarray experiments through a gene set approach is

not trivial because the dynamics of gene expressions inside a gene set can be complex and heterogeneous. This has already been underlined in some of the approaches developed to analyze gene sets [Efron and Tibshirani, 2007; Nueda et al., 2009; Shahbaba et al., 2011; Ackermann and Strimmer, 2009]. Figure 2.3 shows an example of a homogeneous gene set, whereas Figure 2.4 shows an example of a heterogeneous one. Actually, such a heterogeneity is frequently observed [Ackermann and Strimmer, 2009], and cannot be ignored, as genes inside a functional gene set are not expected to change their expression synchronously (Figure 2.5). Moreover this heterogeneity can be biologically meaningful by itself. Prieto et al. [2006] provide an example from a cancer application, where deregulated pathways are of primary biological interest. They identified heterogeneous gene sets linked to acute promyelocytic leukemia. Another example is given by Hu et al.: pathways affected by the HER2, such as the KEGG pathways of 'Ubiquitin mediated proteolysis', 'Glioma', and 'Prostate cancer' were identified by studying heterogeneity [Hu et al., 2013]. The main advantage of detecting the heterogeneity inside a gene set is to detect any change over time whatever the specification of the model for the trends. In other words, the dynamics of gene expression inside a stable gene set will be summarize by a flat slope and no heterogeneity. Hence, in the spirit of [Shahbaba et al., 2011], to find any significant change of the overall expression of genes inside a gene set over time, we suggest to look for any significant trend over time or any heterogeneity between gene trends inside the gene set.

Several approaches have already been proposed to analyze longitudinal measurements of gene expression [Storey et al., 2005; Conesa et al., 2006; Guo et al., 2003; Rajicic et al., 2010; Park et al., 2003; Wolfinger et al., 2001; Luan and Li, 2004], but only a few include gene set analysis [Hummel et al., 2008; Nueda et al., 2009; Wang et al., 2009; Zhang et al., 2011]. Among the latter, all but Nueda et al. [2009] fail to account for possible heterogeneity inside a gene set. An extension of the popular Gene Set Enrichment Analysis (GSEA) method [Subramanian et al., 2005] is available for the analysis of time series data. Unfortunately, it does not account for the structure of longitudinal data, simply treating all observations as independent. The globalANCOVA procedure developed by Hummel et al. [2008] focuses on the comparison of groups, testing whether there is a group influence on change over time of any gene expression inside a gene set. In practice, the global null hypothesis tested is quite flexible relying on the ANOVA framework, but cannot accommodate missing values. Wang et al. [2009] proposed to use a linear mixed effects model to explain gene expression inside a gene set. They considered a random effect for the array level rather than for the patient or the gene level. Zhang et al. [2011] proposed a robust non-parametric approach to compare gene expression dynamics between different treatment-groups. Of note, it is not possible to look at the change
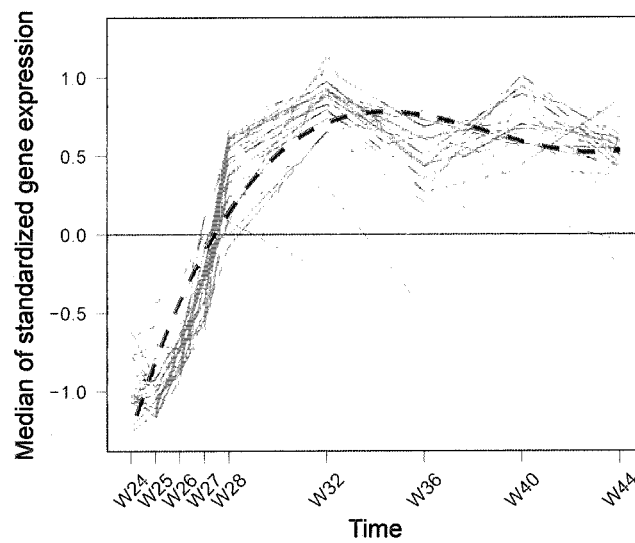
Note: Each line is the median expression of a gene inside this particular gene set across all the patients. The expression of the genes inside this gene set is quite homogeneous and it is easy to identify a global time trend, displayed by the dashed black line (smoothed median). For more information see the presentation of the DALIA-1 trial in section 2.3.3 page 52.

Figure 2.3 – Example of a homogeneous gene set (M1.2: interferon – from the DALIA-1 trial, after treatment interruption)
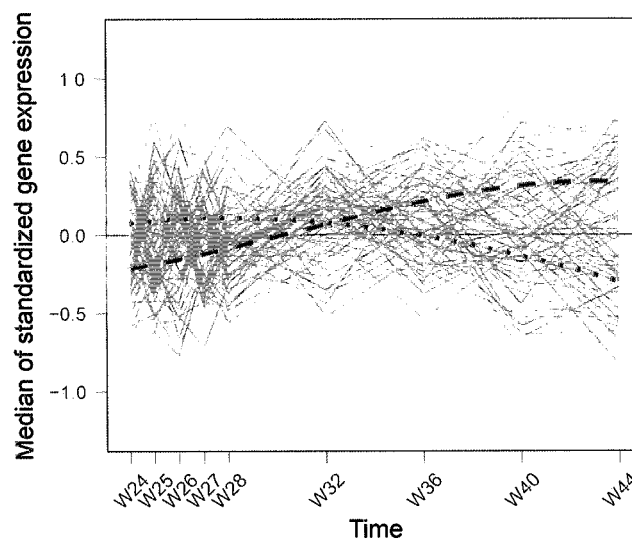
Note: Each line is the median expression of a gene inside this particular gene set across all the patients. The expression of the genes inside this gene set is rather heterogeneous. This makes it difficult to identify any time trend, as the mean expression inside this gene set stays close to zero. However a closer look reveals two distinct time trends, displayed by the two respectively dashed and dotted black lines (smoothed medians). For more information see the presentation of the DALIA-1 trial in section 2.3.3 page 52.

Figure 2.4 – Example of a heterogeneous gene set (M4.16: cell cycle – from the DALIA-1 trial, after treatment interruption)
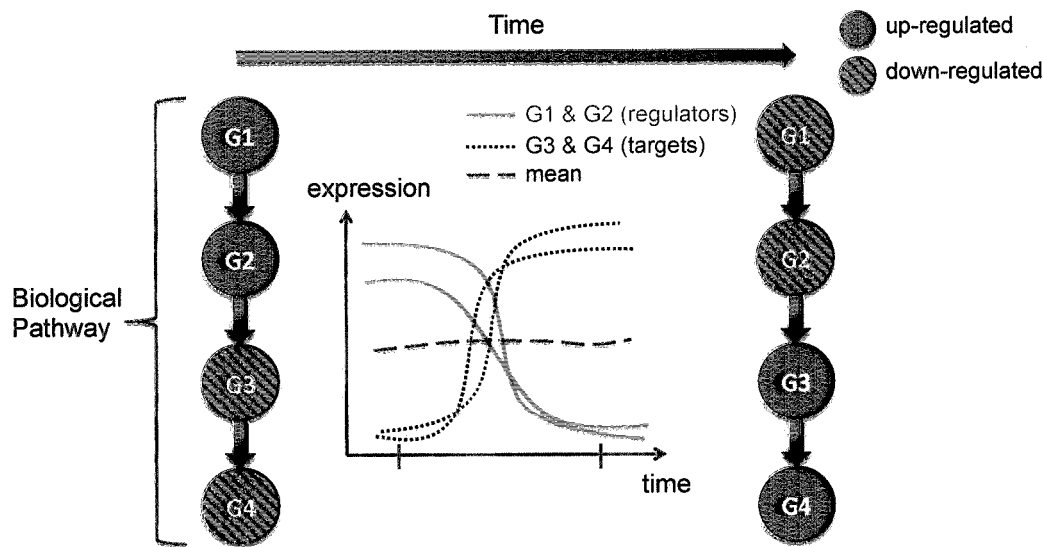
Figure 2.5 – Simplistic representation of heterogeneous gene expression dynamics inside a biological pathway

of gene expression in only one group using either Hummel et al. [2008] or Zhang et al. [2011] approaches. In contrast, the PCA-maSigFun procedure developed by Nueda et al. [2009, 2010] can account for possible heterogeneity inside a gene set. It is based on a Principal Component Analysis (PCA) of each gene set followed by a linear regression of the significant principal components (i.e. components that have a variation above the mean gene variance) over time. However, they did not consider time-course experiments where repeated measures are available for multiple patient. Gene set analysis methods can also be distinguished by their choice of the null hypothesis. Those can be classified into two main types of hypothesis: i) the competitive null hypothesis, that tests the genes inside a given gene set against all the other genes outside the gene set; ii) the self-contained null hypothesis, that only uses the genes inside the gene set of interest [Tian et al., 2005; Goeman and Bühlmann, 2007; Ackermann and Strimmer, 2009]. In the present paper, interest is focused on self-contained null hypotheses because the question was "Which gene sets have a change of gene abundance over time ?".

We propose the implementation of a hypothesis driven method that directly tests the time-course significance of predefined gene sets: the *Time-course Gene Set Analysis* (TcGSA). It relies on the use of linear mixed effect models, a very useful and well established statistical tool [Laird and Ware, 1982; Diggle et al., 2002] especially suited for longitudinal settings. By using all available repeated measures, it provides increased statistical power. TcGSA can accommodate for heterogeneity of gene expression within the gene sets through random effects, and is robust to unbalanced designed due to missing (at random) values thanks to the maximum likelihood estimates. No previously proposed

approach combines all of TcGSA features. A simulation study demonstrated the good statistical performance of the proposed method. It has been applied to two studies: one HIV vaccine trial, and one influenza and pneumoccocal vaccine study [Obermoser et al., 2013], using the same definition of gene sets [Chaussabel et al., 2008] that is increasingly used in systems immunology research [Berry et al., 2010; Zak et al., 2012; Doering et al., 2012; Simonini et al., 2013; Cliff et al., 2013]. Compared to gene-by-gene analyses, TcGSA disclosed changes of additional gene sets that endorse previous conclusions [Obermoser et al., 2013], but also revealed common pathways across the three vaccines.

2.3.2 Time-course Gene Set Analysis method

Time-course Gene Set Analysis

The TcGSA method includes three steps: 1) modeling gene expression in a gene set with mixed models, 2) testing the significance of a gene set, and 3) estimating individual gene profiles.

1. Modeling gene expression in a gene set with mixed models Let S be a gene set of interest. We start by the case of a one group experiment, where each patient act as her/his own respective control, her/his condition changing over time. The expression of genes inside S is modeled over time according to a function f as:

for all the genes $g \in S$,

$$y_{gpi} = \mu + \beta_g + c_{gp} + f_g(t_i) + \varepsilon_{gpi} \tag{1}$$

where y_{gpi} is expression of the gth gene for the pth patient at the ith time, μ is the intercept in the gene set S, β_g is the fixed effect of the gth gene, $c_{gp} \sim \mathcal{N}(0, \sigma_c)$ is a random effect grouped by the gth gene of the pth patient, t_i is the ith measurement time, $\varepsilon_{gpi} \sim \mathcal{N}(0, \sigma)$ is an error term. Finally $f_g(t_i)$ is a function of time, that can be linear, polynomials, etc. Every time coefficient of the trend $f_g(t_i)$ is actually divided into a fixed effect η. (representing the average trend in the gene set S) and a random effect $h_{g,.} \sim \mathcal{N}(0, \sigma_{h.})$ grouped on the gene g, accounting for the possible heterogeneity between the genes in the gene set S. In this paper we focus on three forms for f_g (but other forms, such as exponential, etc. could easily be envisaged):

— linear polynomials:

$$f_g(t) = (\eta_1 + h_{g,1})\,t$$

— cubic polynomials:

$$f_g(t) = (\eta_1 + h_{g,1})\,t + (\eta_2 + h_{g,2})\,t^2 + (\eta_3 + h_{g,3})\,t^3$$

— natural cubic splines:

$$f_g(t) = \sum_{k=1}^{K+1} (\eta_k + h_{g,k}) N_k(t)$$

where the $N_k(t)$ form the natural cubic splines basis [Hastie, 1992] of the time t (with K internal knots), $\eta.$ are the fixed effects of time shared across the gene set S, and $h_g.$ are the random effects of time accounting for possible heterogeneity between genes. $(h_{g,1}, \ldots, h_{g,d}) \sim \mathcal{N}(0, \Sigma_h)$ with d being the degree of the time function, and for $k = 1, \ldots, d$ $h_{g,k} \sim \mathcal{N}(0, \sigma_{h_k})$. Alternatively, one can make the assumption that the patient effect is the same for all the genes. In that case, the random effect c is no longer grouped on the gene level, and the model can be written as:

for all the genes $g \in S$,

$$y_{gpi} = \mu + \beta'_g + c'_p + f_g(t_i) + \varepsilon_{gpi} \tag{1 bis}$$

with $c'_p \sim \mathcal{N}(0, \sigma_{c'})$ the random effect of the patient p, and $\beta'_g \sim \mathcal{N}(0, \sigma_{\beta'})$ the random effect of the gene g. This alternative modeling has the advantage to be more parsimonious than the model (1), with less parameters to be estimated.

Let's now consider the case of a multiple group experiment (such as treatment/vaccine groups for instance). The expression of genes inside S is modeled over time according to a function $f_{g,m}$ that is now stratified on the groups:

for all the genes $g \in S$,

$$y_{mgpi} = \mu + \beta_g + \delta_m + c_{gp} + f_{g,m}(t_i) + \varepsilon_{gpi} \tag{2}$$

where m indicates which group is concerned and δ_m is the fixed intercept of the mth group, everything else being the same as in the model (1).

2. Testing the significance of a gene set

In TcGSA, a "significant" gene set is a gene set whose expression is not stable either over time (in one group experiments) or over groups (in several groups experiments), once between genes and patients variability is taken into account. In other words, we want to test the significance of the time trend while being sensitive to both homogeneous and heterogeneous changes of gene expression over time inside a gene set. Testing the significance of a given gene set S therefore means testing both fixed and random effects at once, in a single test. A likelihood ratio test is the natural way to do so, fitting models under both the null hypothesis and the alternative.

In the case of one group experiment (models (1) and (1 bis)) the null hypothesis (H_0) is that the genes inside S are stable over time, *i.e.* that their expressions are constant and homogeneous over time (all coefficients of the function of time f are not significantly

different from zero). The alternative hypothesis (H_1) is that the genes inside S vary significantly over time:

$$(H_0): \forall k, \eta_k = 0 \quad \text{and} \quad \sigma_{h_k} = 0 \tag{1.0}$$

$$(H_1): \exists k, \eta_k \neq 0 \quad \text{or} \quad \sigma_{h_k} \neq 0 \tag{1.1}$$

In the case of a multiple group experiment (model (2)), the null hypothesis is that inside the gene set S, the evolution of gene expressions over time is the same regardless of the group. The alternative hypothesis is that time trends f are different depending on the group m:

$$(H_0): \forall m, \ f_{g,m}(\cdot) = f_g(\cdot) \tag{2.0}$$

$$(H_1): \exists m, m' \text{ such that } \ f_{g,m}(\cdot) \neq f_{g,m'}(\cdot) \tag{2.1}$$

In both case, one model is fitted under the null hypothesis, and one is fitted under the alternative. The likelihood ratio is then computed.

However, since both fixed and random effects are tested simultaneously in this likelihood ratio, its null distribution is not the standard chi-square distribution (because of boundary constraints due to the variance of random effects). According to Self and Liang [1987], it can be approximated by a mixture distribution of chi-squares with the following formula:

$$LR_{H_0} \sim \sum_{k=q}^{q+r} \binom{r}{k-q} 2^{-r} \chi^2_{(k)}$$

where q is the number of fixed effects and r the number of random effects to be tested simultaneously. This approximation implies that the tested random effects are independent of one another [Stram and Lee, 1994, 1995; Molenberghs and Verbeke, 2007]. This seems an acceptable assumption according to our simulations under the null hypothesis. See Figures 2.6 and 2.7 which compare 100,000 LRTs computed on simulated gene sets (with similar settings as those of the DALIA-1 trial presented in section 2.3.3 page 52) under the null hypothesis (no effect of time, either as a fixed effect or a random effect, using a cubic polynomial function of time): even though the random effects (three functions of time) are not independent, the approximation seems quite valid. This allows to compute a p-value for the significance of the variation of a given gene set over time.

When several gene sets are investigated at a time, it is necessary to take into account multiple testing. A number of procedures are available to do so [Dudoit and Van der Laan, 2008]. As the TcGSA is mostly an exploratory analysis (even though hypothesis driven in the sense that gene set are defined *a priori*), we recommend using the Benjamini-Yekutieli procedure for controlling the False Discovery Rate [Yekutieli and Benjamini, 2001], as gene sets are necessarily correlated between each others and this procedure is robust to some
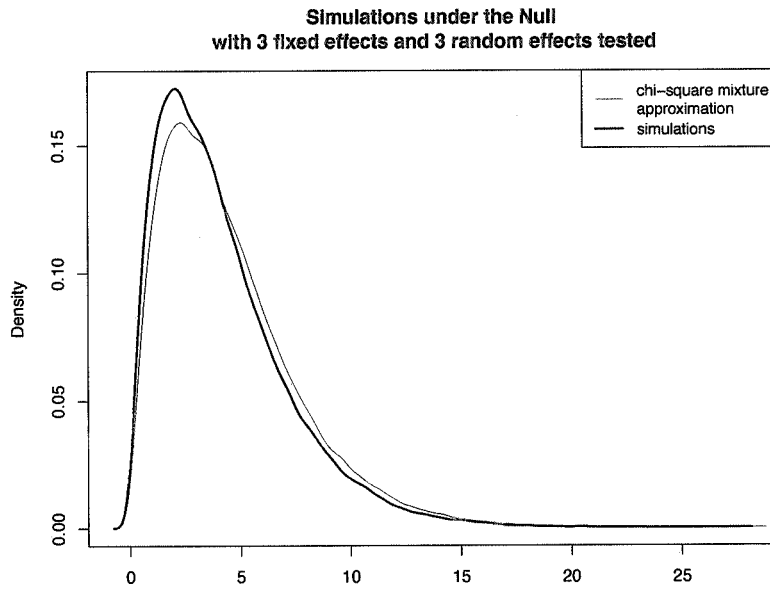
Figure 2.6 – Density plot for both the 100,000 simulations under the null and a 100,000 sample of the corresponding χ^2 mixture approximation.
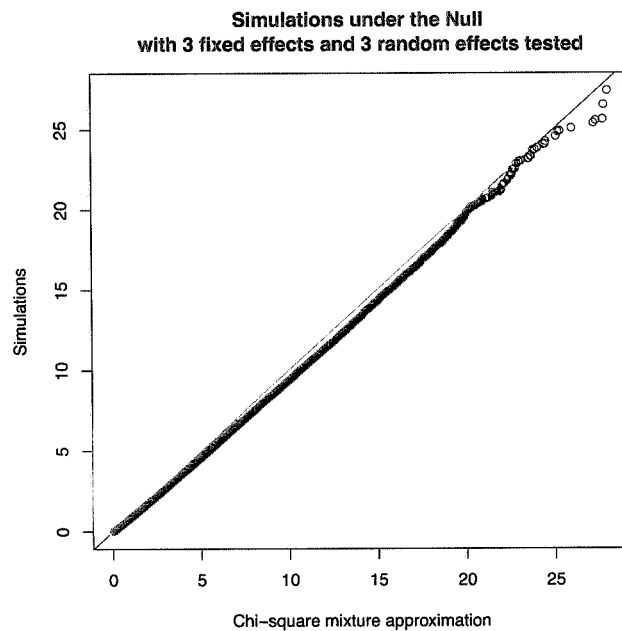
Figure 2.7 – Quantile-Quantile plot comparing the 100,000 simulations under the null to a 100,000 sample of the corresponding χ^2 mixture approximation.

of these dependances. Other mutliple testing correction procedures are available in the *TcGSA* ℛ package.

3. Estimation of individual gene profiles In the estimation of linear mixed model, it is common to use the Restricted Maximum Likelihood (REML) instead of the classic Maximum Likelihood (ML) in order to avoid biased estimates of the variance components [Harville, 1977]. But note that REML cannot be used to estimate the likelihood ratios presented here. Indeed, REML estimation of the likelihood ratio between two models can only be used when both models have the same fixed part [Snijders and Bosker, 2012]. Since here the compared models (under H_0 and under H_1) have different fixed components (due to the η. coefficients under H_1), the use of ML estimation is needed.

For the inference of the random effects, Best Linear Unbiased Predictor (BLUP) are used [Verbeke and Molenberghs, 2000], giving access to estimations of a single profile for each gene among a gene set, in each patient (Figures 2.9 and 2.10 where the medians of those profiles over the patients are represented in our motivating example). As a result, the estimations from the mixed model are shrunken towards the average expression inside the gene set. This shrinkage occurs when the residuals variability is relatively large compared to the the random effects estimated variances [Verbeke and Molenberghs, 2000]. The mixed model uses the repeated pattern of the longitudinal measurements to structure the variation. Its estimations give smoother trajectories for the genes than the raw data, which makes the general evolution of the set clearer [Hitchcock et al., 2007], as it can be seen in Figures 2.9 and 2.10.

Characterization and visualization of Dynamics

Dynamic of a significant gene set Once a gene set S has been identified as significant, a summary of its dynamic over time is needed. However, due to the possible heterogeneity of S, giving a summary representation of S dynamic is not obvious. We propose to automatically identify the number of trends in a significant gene set. Predicted gene expressions are clustered, and the optimal number of trends is selected with the gap statistic [Tibshirani et al., 2001]. It is a formalization of the elbow criterion for the within-cluster variance. In order to determine the optimal partition of each gene set here, the gap statistics is applied onto a hierarchical clustering of gene expressions inside each gene set. Then the median within each of the identified clusters can summarize each trend. Therefore, gene sets are actually split when heterogeneous, before being summarized. The predicted gene expression is used (and not the observed expression) because smoothness of trajectories facilitates classification [Hitchcock et al., 2007]. Examples of such representations are given in Figures 2.9 and 2.10.

Global dynamics Most often, TcGSA will be used to investigate a large number of gene sets (from a few hundreds to a few thousands). This multiplicity can make visualization of the results more challenging, in addition of requiring a multiple testing correction. TcGSA is designed to identify gene sets that shows a simultaneous evolution of gene expression, but possibly of a small intensity. The method can therefore be quite sensitive, and it can be of interest to rank the significant gene sets to identify the most acute signals. The likelihood ratio provides insight on the magnitude of the variation of each gene set. The percentile of their corresponding likelihood ratio gives an idea of the importance of the variation for a significant gene set. Examples of such representations are given in Figure 2.11.

2.3.3 Applications of Time-course Gene Set Analysis

Motivating example: the DALIA-1 trial

The DALIA-1 trial is a phase 1 therapeutic HIV vaccine trial whose details are described on http://clinicaltrials.gov (ClinicalTrial.gov identifier: NCT00796770) and in Lévy et al. [2014]. The vaccine candidate was based on ex-vivo generated interferon-α dendritic cells loaded with HIV-1 lipopeptides and activated with lipopolysaccharide. The objectives of the trial was to evaluate the safety of the strategy and to evaluate the immune response to the vaccine. For the purpose of the present paper, we focus on the gene expression component of this study. Gene abundance in whole blood was measured through Illumina® HumanHT-12 v4 Expression BeadChips.

The DALIA-1 trial design All of the nineteen HIV infected patients received the therapeutic vaccine while under antiretroviral treatment. The patients received four injections at week 0, 4, 8 and 12. This vaccination period was followed by an antiretroviral treatment interruption (ATI) at week 24. The patients were followed up to week 48. Antiretroviral treatment was resumed from week 24 to week 48 at any time under the following criteria: i) if the patients or their doctors wished so; ii) if CD4+ T-cell count was <350 cells/μL and <25% of total lymphocytes. Fourteen time points (five in pre-ATI from week 0 to week 22, and nine in post-ATI from week 24 to week 44) were used for this analysis (Figure 2.8). One patient was removed from the analysis as his/her antiretroviral treatment compliance was irregular during the vaccination phase.

In the following analysis, two distinct datasets were considered: pre-ATI and post-ATI. The two datasets were normalized separately – via a normal-exponential convolution model [Xie et al., 2009; Shi et al., 2010], followed by the application of the *ComBat* method [Johnson et al., 2007] to correct for batch effects. Splitting the data allows us to study sep-
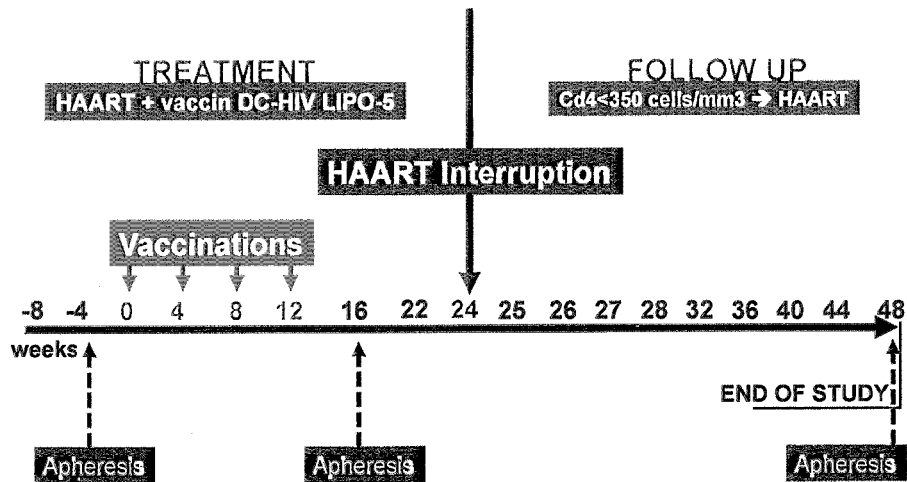
Note: Gene expression was measured at each time point, represented by a week number above the time axis. The trial was composed of two separated stages: (1) the treatment phase, during which the patients were vaccinated but remained under antiretroviral treatment; and(2) the follow- up phase commencing after the week 24 antiretroviral treatment interruption. Those two phases will be referred as pre-ATI and post-ATI respectively. The three apheresis time points were removed from the analysis due to a possible effect of the apheresis on the gene expression samples, and so was the first measurement (week -8) occurring right at the inclusion in the study.

Figure 2.8 – DALIA-1 trial design

arately the vaccine effect and the treatment interruption, otherwise the ATI effect would mask any noticeable vaccine effect, because of the huge modification of gene expression related to viral replication [Bécavin et al., 2011; Bosinger et al., 2012]. We investigated the gene sets defined by Chaussabel et al. [2008], which are oriented towards the immune system. The definition and annotations of those 260 gene sets (called 'Modules') are available online (http://www.biir.net/public_wikis/module_annotation/V2_Trial_8_Modules).

Pre-ATI: the vaccination phase During the vaccination period, a standard gene-by-gene mixed model analysis did not found any significant change of gene abundance at a 5% False Discovery Rate (Table 2.1). However, during this vaccination phase, cytokines production analysis of the same blood samples (as measured by Luminex or intracellular staining) have showed that a response was induced by the vaccine at week 16 [Lévy et al., 2014]. Therefore, one expected to observe a signal at the gene expression level between week 0 and week 16, the gene expression preceding molecular activation. Although the measurements were not performed in the hours or days following vaccination, the changes of gene abundance may reflect a change of the equilibrium of the overall expression in some gene sets. This kind of results has already been reported in cross-sectional studies [Murohashi et al., 2010]. Likewise, GSEA for time-series did not identify any significant

gene set during vaccination. This can be explained by the lack of power of GSEA for time-series, as this method does not take into account the repeated structure of the data and is not suitable for longitudinal measurements.

We applied the Time Course Gene set Analysis (Methods) that allows to detect any change over time of gene abundance inside a gene set by detecting either trends over time or heterogeneity between gene dynamics. Fitting the model (1) with a cubic spline function of time, 69 gene sets out of 260 turned out to vary significantly. Figure 2.9 displays the raw and estimated gene expressions of 3 of the significant gene sets identified by TcGSA: T-cell, inflammation and B-cell gene sets. The identification of gene sets such as M4.1: T-cell (that includes CD402, CCR7, BCl2) was expected with regards to the CD4 T-cell response observed at Week 16 [Lévy et al., 2014]. Also, the gene sets M4.6: inflammation and M6.7: B-cell are good examples of how smoothing from the estimations can give a much clearer dynamic pattern compared to the raw expression (Methods).

Post-ATI: after antiretroviral treatment interruption   Model (1) was then fitted to the data after antiretroviral treatment interruption that occurred at week 24: 216 gene sets out of 260 were found to be significant. Figure 2.10 displays the raw and estimated expressions of nine of those significant gene sets. It features heterogeneous gene sets, such as M4.16 and M7.1, which are both also good examples of the shrinkage that occurs with the estimations (Methods).

The large number of significant gene sets post-ATI illustrates the tremendous impact of the treatment interruption on the organism. Followed by a viral rebound, the treatment interruption is indeed a major event that triggers the expression of thousand of genes. Indeed, a gene-by-gene analysis revealed 7,534 significant probes (more than 20% of the investigated probes – an unusually high number of differentially expressed genes). The immune system is very much in demand during the viral rebound. Therefore most of the gene sets from the Modules defined by Chaussabel et al. [2008] are activated, as they are tightly linked with the immune system activity. Of particular interest are the three gene sets M1.2, M3.4 and M5.12 which are all annotated as *interferon*-related. These three gene sets exhibit similar dynamics (Figure 2.10). Such a timely upregulation was expected, as it is linked to the viral rebound after treatment interruption and was previously reported [Bécavin et al., 2011; Bosinger et al., 2012]. The gene set M3.4 is also linked with *antiviral response*.

Another application: influenza and pneumococcal vaccines responses

In a recent paper, Obermoser et al. [2013] investigated the response to influenza and pneumococcal vaccines in healthy individuals at the gene expression level.

	Pre-ATI	Post-ATI	Units
Gene-by-gene	0	7,534	probes [a]
GSEA for time series	0	67	genes set [b]
TcGSA linear	23	203	gene sets [b]
TcGSA cubic	69	216	gene sets [b]
TcGSA splines	68	219	gene sets [b]

[a] 32,978 probes investigated after filtering

[b] 260 immune-related gene sets investigated (29 gene sets were automatically discarded because less than 10 probes were observed)

Table 2.1 – Number of significant units in DALIA-1 at a FDR of 5%
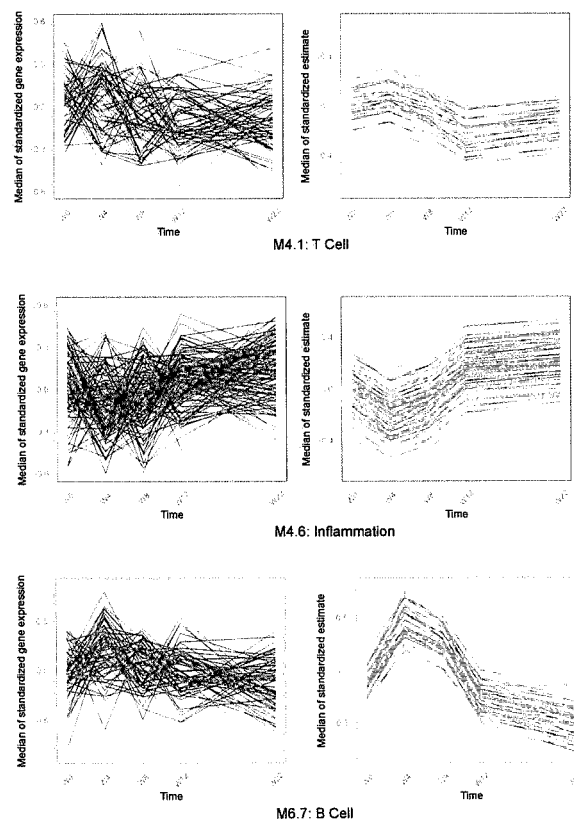
M4.1: T Cell

M4.6: Inflammation

M6.7: B Cell

Note: Each line is the median over the patients of the expression of one gene. Each graph shows all the genes in one particular gene set. The left graph displays the raw gene expression, the right one displays the estimations from the mixed model for the same gene set. The expressions have been centered and reduced for this representation. The percentile likelihood ratio associated with the displayed gene sets is given as an indication of their rank of significance.

Figure 2.9 – Four significant gene sets during pre-ATI in DALIA-1
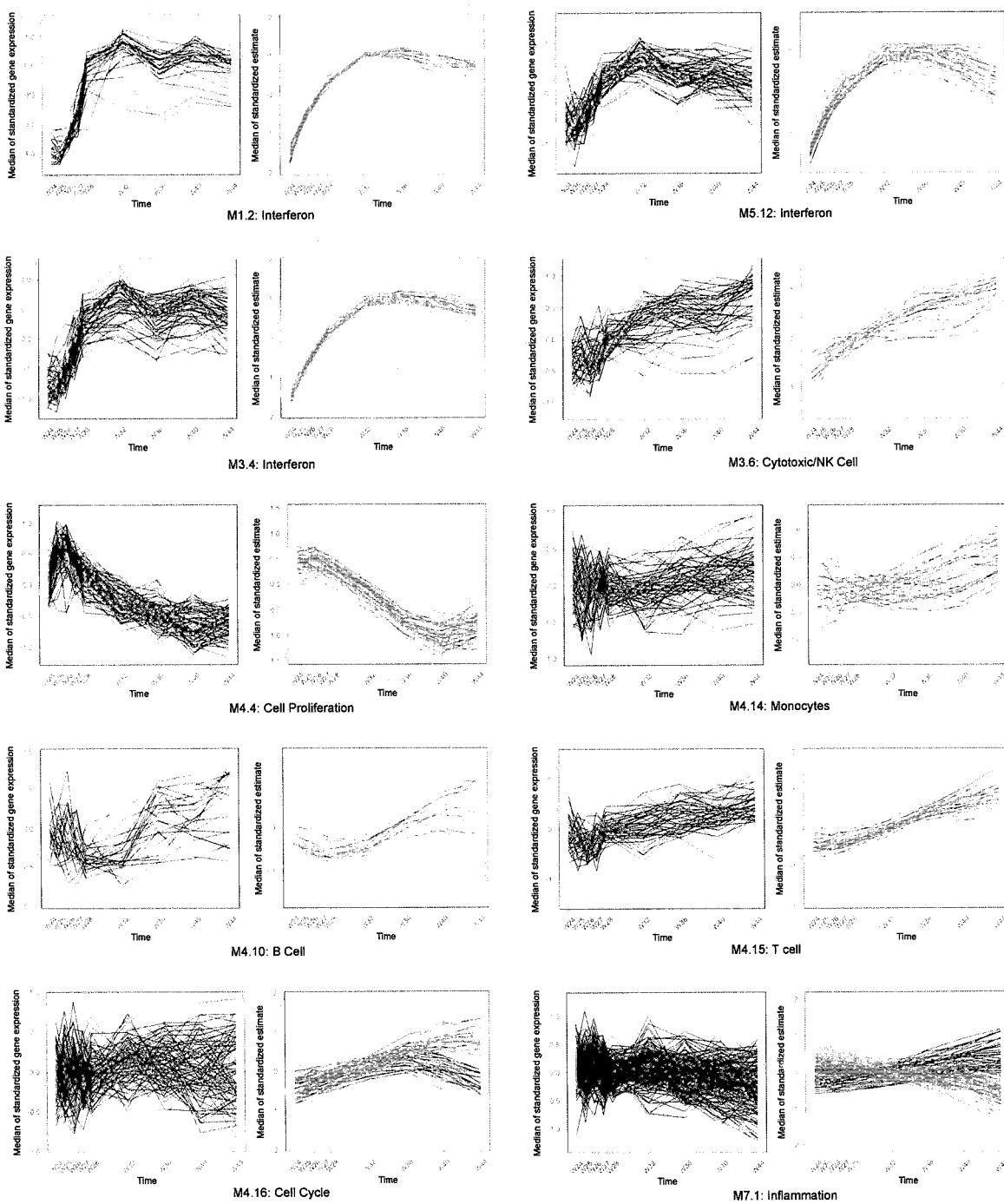
Note: Each line is the median over the patients of the expression of one gene. Each graph shows all the genes in one particular gene set. The left graph displays the raw gene expression, the right one displays the estimations from the mixed model for the same gene set. If several dynamics are identified by the gap statistics among the estimated expressions inside one gene set, they are displayed in different colors – such as for the gene sets M 4.16 and M 7.1 that each features three different dynamics. The expressions have been centered and reduced for this representation. The percentile likelihood ratio associated with the displayed gene sets gives an indication on the relative importance of their variation.

Figure 2.10 – Ten significant gene sets during post-ATI in DALIA-1

Study design Healthy, young adults were randomly split in three groups of six volunteers each, receiving either a 2009-2010 seasonal influenza vaccine (Fluzone), a 23-valent pneumococcal vaccine (Pneumovax23), or a placebo (saline injections). Blood samples were collected at days -7, 0, 1, 3, 7, 10, 14, 21, and 28 to measure gene expression in whole blood. A more detailed description of the study can be found in Obermoser et al. [2013].

Original analysis In their modular analysis, Obermoser et al. [2013] focused on 62 of the 260 available gene sets defined in Chaussabel et al. [2008]. They investigated the changes of gene expression in those 62 gene sets for each of the seven time points from day 1 to day 28 in regards of the baseline, that was considered as the average of the two measurements at days -7 and 0. So hierarchical structure of the data was not taken into account. The three arms (saline, flu and pneumococcal) were analyzed separately, and only significant gene sets at day 1 and day 7 (not further on) are presented in their paper. Changes in eight gene sets were common to both vaccines: M4.6 (inflammation), M6.6 and M6.13 (apoptosis/cell death) and modules M4.1 and M4.15 (T cells), M4.3 (protein synthesis), M5.11, and M6.9 (no functional annotation). Nine gene sets were uniquely changing after the influenza vaccine, three were associated with antiviral responses (M1.2, M3.4, M5.12) and included genes coding for interferon (IFN)-inducible molecules. Six gene sets were uniquely responsive to the pneumococcal vaccine. Of these, five were modules including genes associated with inflammation: M3.2, M4.2, M4.13, M5.1 and M5.7.

TcGSA results To compare the gene expression at the gene set level between the vaccine arm (flu or pneumococcal) and the placebo (saline) arm, we applied TcGSA on these data using model (2) (for each vaccine separately). In both vaccines, a large response is observed at Day 1. To avoid smoothing down the expression at $t_i = 1$, we used the following function of time to model the dynamic evolution of gene expression:

$$f_m(t_i) = (\eta_m + h_g)\mathbb{1}_{\{t_i=1\}} + (\eta'_m + h'_g)\mathbb{1}_{\{t_i \neq 1\}}$$

with $(h_g, h'_g) \sim \mathcal{N}(0, \sigma_h)$, and m the group (either vaccine or placebo).

Most of the 62 investigated gene sets presented a significantly different evolution in vaccine arms compared to the placebo arm. Globally, the intensity of the response was stronger with the pneumoccocal vaccine than with the flu vaccine (Figure 2.11). The early response induced by the pneumococcal vaccine was dominated by inflammation whereas the top signal triggered by the flu vaccine involved an interferon signature (Figure 2.11B and 2.11D). In both vaccine, a T-cell response was also visible. In the pneumoccocal vaccine, a plasma cell signal, in association with cell cycle gene sets (Figure 2.11A and

2.11C), started at Day 7 until Day 14. This plasma blast signal was much less clear in the flu vaccine (Figure 2.11B and 2.11D). This is in agreement with the results of Obermoser et al. modular analysis.

TcGSA offers an extended and appropriate hierarchical analysis of these data. It provides a truly longitudinal insight into the vaccine responses, that are intrinsically compared to the placebo response. One of the main difference from the results presented in Obermoser et al. [2013] is that, according to our analysis, the inflammation gene sets (M3.2, M4.13, M5.1 and M5.7) were also involved with the flu vaccine and were not specific to the pneumoccocal vaccine. This result is important as it means that both vaccine involved these inflammatory pathways. This result was not obvious from the original analysis because their approach was less powerful compared to the TcGSA.

Assessment of statistical performances on simulated data

In order to assess the behavior of the proposed method, a simulation study of TcGSA has been performed. The simulation scheme was chosen to be very close to the motivating exemple: the DALIA-1 trial. In each simulation run, gene expression data was simulated for 20 patients at 8 time points. 5,000 genes were simulated, divided into 100 non overlapping gene sets of 50 genes each. Each of the 100 gene sets was either simulated under (H_0) or (H_1). The proportion of genes under (H_1) varied between 0%, 27% (which corresponds to results found in pre-ATI) and 85% (which corresponds to results found in post-ATI). When there are gene sets simulated under (H_1), 75% of those were homogeneous (simulated with parameters close to those estimated for gene set M1.2 in DALIA post-ATI – see Figure 2.10) while the remaining 25% were heterogeneous (simulated with parameters close to those estimated for gene set M7.1 in DALIA post-ATI – see Figure 2.10).

Statistical performances of the proposed method are presented in Table 2.2. Without correcting for the fact that 100 gene set were investigated by TcGSA at each simulation runs, the average Type I error (the probability of rejecting H_0 given that H_0 is actually true) over a hundred runs was between 0.03 and 0.07 depending on the situation. But as soon as a control of the FDR was used, the Type-I error rate dropped well below 1%, regardless of the flexibility of the time function estimated (linear or cubic polynomials). The average statistical power (the probability of rejecting H_0 given that H_0 is actually false) is very good, always above 0.8 (dropping a little bit after multiple testing correction as expected).

Two other methods were also evaluated on those simulations, namely globalANCOVA [Hummel et al., 2008] (using either permutations or an approximation to compute p-values) and GSEA for time series. Their statistical performances are also presented in Table 2.2. Type I error is always well controlled for both those methods. However,
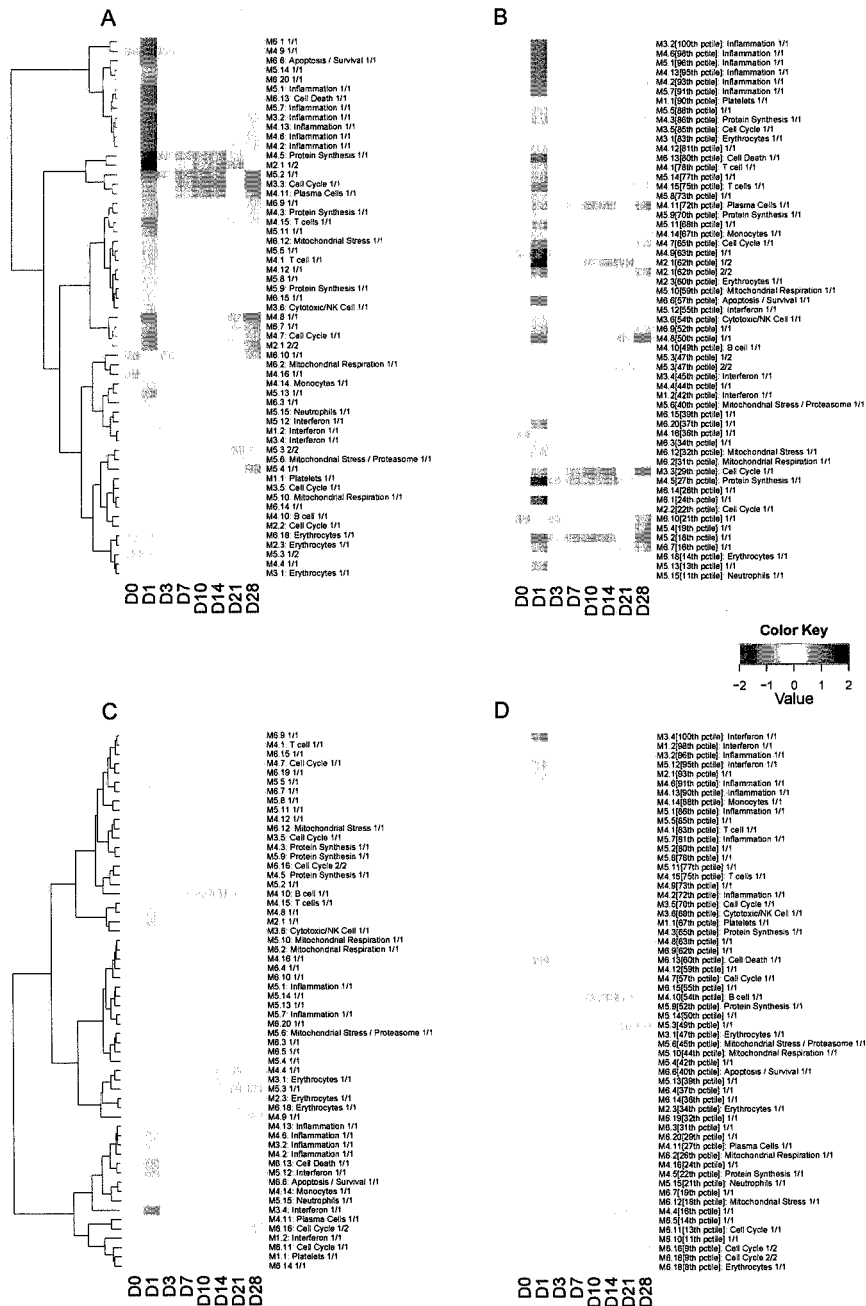
Note: The median estimated gene expression over the patients is used for each trend. Each trend has seen its values reduced (so that its variance is 1) in order to make the dynamics more comparable. Each row is a group of gene having the same trend inside a gene set, and each column is a time point. The color key represents the median of the standardized estimation of gene expression over the patients for a given trend in a significant gene set. It is becomes red as median expression is up-regulated or blue as it is down-regulated compared to the value in the placebo (saline) at the same time. A and C show the hierarchically clustered trends for pneumoccocal and flu respectively. B and D show the same trends but instead ranked by decreasing likelihood ratio percentiles of the associated gene set, for pneumoccocal and flu respectively.

Figure 2.11 – Heatmap of estimated dynamics from the significant gene sets among the 62 investigated gene sets when comparing vaccines arms to placebo arm

Percentage of simulated gene sets under H_1	Method	Type I error	Type I error after MTC*	Statistical power	Statistical power after MTC*
0%	TcGSA (linear)	0.0394	0.0002	-	-
0%	TcGSA (cubic)	0.0649	0.0004	-	-
0%	globalANCOVA (perm)	0.0483	0.0001	-	-
0%	globalANCOVA (approx)	0.0006	0	-	-
0%	GSEA for time series	0	0	-	-
27%	TcGSA (linear)	-	-	0.883	0.829
27%	TcGSA (cubic)	-	-	0.882	0.810
27%	globalANCOVA (perm)	-	-	0.787	0.706
27%	globalANCOVA (approx)	-	-	0.660	0.510
27%	GSEA for time series	-	-	0.459	0.214
85%	TcGSA (linear)	-	-	0.885	0.847
85%	TcGSA (cubic)	-	-	0.882	0.833
85%	globalANCOVA (perm)	-	-	0.785	0.728
85%	globalANCOVA (approx)	-	-	0.660	0.549
85%	GSEA for time series	-	-	0.289	0.074

* Multiple Testing Correction: performed via Benjamini-Yekutieli procedure with a 5% threshold.

Note: In each simulation, 100 gene sets are simulated and significance level $\alpha = 5\%$ is applied. This table displays the Type I error and the statistical power means over a hundred simulation runs for 3 different situations (0%, 27% and 85% of simulated gene sets are simulated under H_1). Whenever the percentage of gene sets simulated under H_1 is not null, 25% of the gene sets simulated under H_1 are heterogeneous, the remaining 75% being homogeneous. Type I error is the probability of rejecting H_0 given that H_0 is true, i.e. for declaring a gene set significant when it actually is not. Statistical power is the probability of rejecting H_0 given that H_1 is true, i.e. for declaring a gene set significant when it actually is. Three methods are evaluated: i) TcGSA, the proposed approach, fitted either with a linear or with a cubic function of time ; ii) the GlobalANCOVA procedure [Hummel et al., 2008] in which p-values are either computed by permutation (10,000) or approximated; iii) the GSEA for time series [Subramanian et al., 2005]. Default values are used for the various methods.

Table 2.2 – Assessment of statistical performances through a simulation study

GSEA for time series has very low statistical power (as low as 10 times less than TcGSA after multiple testing correction when there is a high proportion of significant gene sets). globalANCOVA, whose global null hypothesis is not so different from the one tested in TcGSA, performs quite well in terms of statistical power. Nonetheless it is still about 10% below TcGSA performances.

Those simulation results confirm that the higher number of selected gene sets by TcGSA in the two real-life examples presented in this paper are mainly due to the increased power of gene set analysis over gene-by-gene analysis (when repeated structure of the measurement is properly accounted for), and not to a large number of false positives.

2.3.4   Discussion of Time-course Gene Set Analysis

In this paper, we present a method to analyze repeated measurements of gene expression using a gene set approach. Provided that the definition of the gene sets is relevant, this method helps with detecting and interpreting subtle changes of gene expression over time. In our applications where the same definition of gene sets has been applied, we were able to compare the response to several vaccines (against HIV, Influenza and Pneumococcus). Interestingly, we found common pathways that were triggered by all three vaccines, mostly related to inflammation, as well as pathways specific to each vaccine.

The capacity of the proposed approach to detect subtle changes of gene expression is due to two main factors: i) the use of a predefined gene sets that are functionally related ii) the use of all available information, taking advantage of repeated measurements using mixed models. Measurements of gene expression data in longitudinal studies may be missing because of missed visits or poor quality of the samples, leading to unbalanced data. Missing at random (MAR) processes (i.e. when the probability of missing data is associated to the previously measured information) may lead to biased estimates when using least squares or generalized estimating equations [Diggle et al., 2002]. TcGSA can cope with such issues because of the use of Maximum Likelihood to estimate the parameters of the mixed models. This is an advantage of the TcGSA approach over those of Hummel et al. [2008] or Nueda et al. [2010].

An increasing number of gene sets databases are available: KEGG [Kanehisa and Goto, 2000], Gene Ontology [Ashburner et al., 2000], Modules [Chaussabel et al., 2008]. An immune related subset of Gene Ontology as well as an immune related subset of KEGG pathway have been used in additional analyses (Appendix B page 125). The choice of the database used for the analysis impacts the interpretability but also the limitations of TcGSA. The efficiency of TcGSA will vary according to the number of genes represented in each gene set. The size of a given gene set has an impact on its significance, as the more

genes it includes, the more likely a significant variation will be detected. The average size of the Chaussabel's V2 modules is 55 genes. 17% of the 260 modules include more than 100 genes, and 31% less than 20 genes. For small gene sets, the normality assumptions of random effects of the models (1), (1bis) and (2) are questionable. Nevertheless, even though we expect that the models could be miss-specified in many cases (if not all), the objective of such an analysis is to detect any significant variation over time (in the spirit of Shahbaba et al. [2011], a significant variability of the trajectories between the genes inside a gene set indicates a change over time regardless of the fixed effects specification). The use of flexible time functions may help to get a better fit of dynamics although beyond cubic polynomials it did not have a substantial impact in our motivating example – see Table 2.1. These results vary according to the dataset and the number of time points available, and we recommend to try several models to check the robustness of the results.

Several extensions of the TcGSA are possible for its use in other contexts. One can also model time trends with a random effect grouped on the patient level as of $\gamma_{p,\cdot}$, instead of on the gene level as in models (1), (1bis) and (2). This identifies gene sets whose dynamic differs across the patients. This option is also implemented in the *TcGSA* ® package. TcGSA could easily be adapted to mRNA counts data. In that case, generalized linear mixed effects models should be used, with a Poisson distribution for instance, instead of linear mixed effect models that rely on a Gaussian assumption.

In conclusion, the method presented gives a solution for the full exploitation of any repeated measurements of gene expression data based on a gene set analysis where a great sensibility to detect subtle change, while controlling false discovery, is needed.

Implementation

The TcGSA method has been implemented in ® as a package TcGSA, whose latest release is available from the CRAN repository (http://cran.r-project.org/web/packages/TcGSA/index.html).

3 Integrative analyses of gene expression data in two vaccine trials

Abstract:

This chapter presents two integrative analyses applied to vaccine trials. The first application concerned the HIV therapeutic trial DALIA-1 (presented in the previous chapter). The second example was about the differential effect of a trivalent influenza vaccine according to sex. Both analyses try to unravel some underlying biological mechanisms triggered by the vaccine, in order to understand why some patients have a better immune response following vaccination than others. In the DALIA-1 trial, the integration of gene sets for which the expression changed over time and relevant immune variables revealed that lower expression of inflammatory pathway during vaccination was associated with a better immune response after antiretroviral treatment interruption. In the other application on the flu vaccine, a high-dimensional analysis of interaction identified a group of genes linked to lipid metabolism that could mediate the sex effect on the antibody response following vaccination. In particular, further analysis showed that this mediation was actually associated with circulating testosterone levels.

Key Words: Immuno-endocrine; Integrative analysis; Sexual dimorphism; Sparse Partial Least Squares; Supervised analysis

Contents

Valorisation: section 3.2 is mainly part of an article written in collaboration with David Furman that was published in *PNAS* (The original article is provided in Appendix E page 143):

3.1 Integrative analysis of the DALIA-1 trial

The DALIA-1 trial, as described in subsection 2.3.3 page 52, is a phase-I therapeutic HIV vaccine trial. Its primary goal was to evaluate the safety of the vaccine, and also the immune response associated with it as a secondary endpoint. During the vaccination phase, TcGSA identified 69 Modules whose expression varied significantly over time, encompassing a total of 5,399 probes. In a system biology perspective, it is now of interest to relate these gene expression to other measurements of the immune response to the vaccine.

3.1.1 Immune measurements in the DALIA-1 trial

During the DALIA-1 trial, in addition to the repeated measurements of gene expression, various immune markers were measured periodically in 16 patients from in vitro T cell assays. Immune-related cytokines measured by the Multiplex technology and cell polyfunctionality measured via intra-cellular staining summarized through Immunogenicity scores [Lévy et al., 2014] were available before and after vaccination (at weeks 0 and 16, see Figure 2.8).

One fundamental hypothesis of the trial was the impact of the vaccine response on the dynamics of viral load after treatment interruption. Figure 3.1 shows the multimodal non-parametric density estimation of the observed peak viral load, suggesting that there might be good responders (log(peak viral load) $\leq$ 5) and bad responders (log(peak viral load) $>$ 5). According to Lévy et al. [2014] log(peak viral load) $=$ 5 could be a good threshold to distinguish between good and bad responders. The immune responses (measured at week 16) most associated with this peak viral load are the following seven variables: the cytokines luminex score, the T-helper 1 score, the CD4 cell polyfunctionnality, the interleukins 2, 13 and 21 productions, and the interferon γ production. We investigated the potential associations between the vaccine elicited gene expression (at week 16) on the one hand, and those immunological responses as well as the viral dynamics after treatment interruption on the other hand. This may help to understand the mechanism and the determinants of the vaccine effect.

3.1.2 Sparse Partial Least Squares method

Sparse Partial Least Squares (sPLS) is a penalized dimension reduction method that explores two data sets at once, constructing a multivariate linear model. It is build on the Partial Least Squares (PLS) approach, onto which a LASSO kind of penalization is added. Appendix C page 131 gives an overview of this method. The sPLS relates two
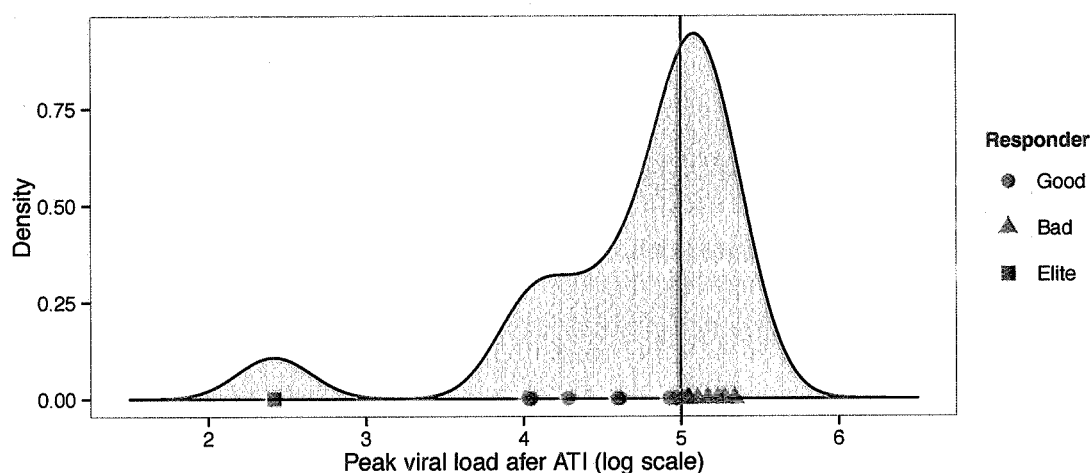
Note: The density of the logarithm transformation peak viral load appears multimodal. Median is 5.

Figure 3.1 – Distribution of patients peak viral load after antiretroviral treatment interruption in the DALIA-1 trial

data matrix X and Y by seeking to iteratively construct sparse linear combinations of their respective columns, so that those respective combinations maximize the covariance between themselves.

3.1.3 Supervised multivariate integrative analysis results

Let X be a 16 × 5399 matrix and Y be a 16 × 8 matrix. Both matrix contains data from the 16 patients that were sampled fro immunological measurements at week 16. X contains the measured expression at week 16 for the 5399 probes that were measured and that participate in the 69 significant Modules according to TcGSA results (section 2.3.3). Y contains the data for the following eight variables measured at 16 weeks:

— interferon-γ production
— interleukin 2 production
— interleukin 13 production
— interleukin 21 production
— CD4 cells polyfunctionnality [Lévy et al., 2014]
— T-helper score [Lévy et al., 2014]
— Luminex score [Lévy et al., 2014]

and the maximum observed viral load after antiretroviral treatment interruption (ATI):

— log(peak VL) > 5 (post ATI)

Both X and Y components were chosen sparse. Four components were kept for the analysis, mainly because this is the minimum number for all Y to be selected at least once.

Yet this was sufficient for explaining a lot of the immune variability in Y (Figure 3.2A). A sequential cross-validated (leave-one-out) optimization of the mean R^2 between the Y variables selected by a given component explained by the corresponding X component gives the following optimum: for X 24, 13, 8 and 1 genes were respectively selected for each component (Table 3.1) ; for Y 2, 2, 3 and 4 variables were respectively selected for each component. Figure 3.2B shows the importance of each immune variable on each of the four Y components. The maximum viral load (log(peak VL) >5) was selected only by the third Y component. Therefore, the matching third X component separated very well the patients depending on their peak viral load level (Figure 3.3).

Figure 3.4 displays the strength of the association between genes that are selected on any of the four components and the immune variables. The log(peak viral load)>5 has an opposite profile compared the other immune variables. This makes sense: the higher the cytokines productions (i.e. the better the immune response), the lower the viral rebound. In addition, a clear inflammatory response was positively correlated with a bad response to the vaccine and lower cytokine productions. Each selected gene was annotated with the module it belongs to and the component was selected in. One striking result was that many selected Modules were not annotated.

Robustness analyses

A number of robustness analyses were performed: other approaches were tried out (univariate response LASSO model, Variable Selection Using Random Forests – VSURF [Genuer et al., 2010]), as well as other penalty parameters and robustness to outlier samples. The univariate LASSO models were hard to interpret because of frequent difficulties in the tuning of the penalty parameters. The VSURF models have the advantage of considering non linear relations between their univariate outcome and the explanatory variables from X, but the disadvantage to be limited to a univariate response variable. A 100 bootstrap (equiprobable sampling with replacement) samples of the individuals were created. On each of these samples, the sPLS has been run with 4 components. The keepX and keepY numbers (the numbers of genes and of response variables respectively to be kept for each of the 4 components) were automatically determined, following the same principle as the main analysis: the automatic procedure maximizes the mean cross-validated (leave-one-out) R^2 of the Y variables selected for each component sequentially (above and Appendix C page 131). The resulting sPLS models are not comparable component by component, as the selected Y variables on each component are different from one model to another. Therefore, interest is focused on wether a module was selected in one of the 4 components or not. Figure 3.5 and Table 3.2 display the results (only the modules that were selected in at least in 1 of the 100 bootstraped models are featured).
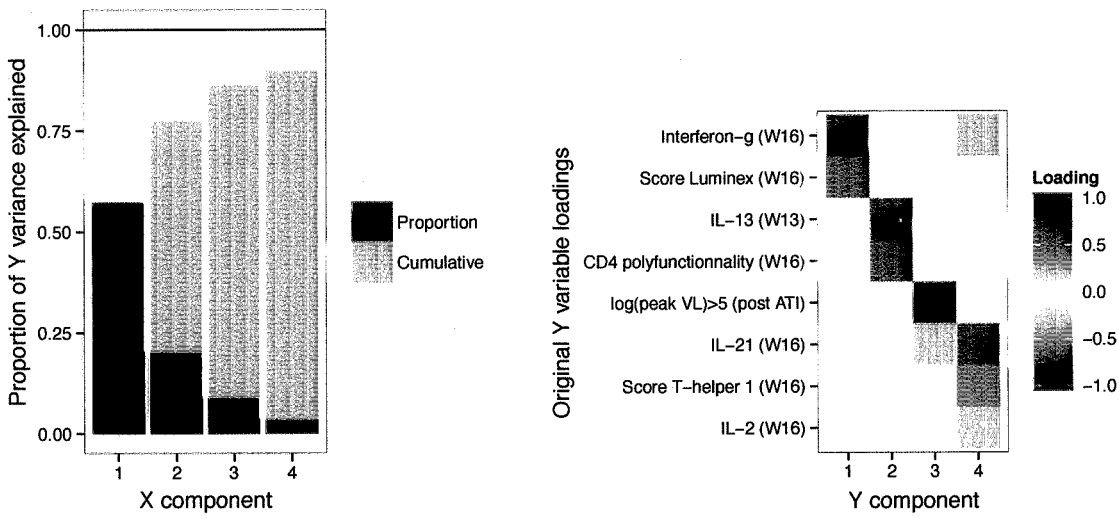
A: Proportion of immune phenotype variability (Y) explained by gene expression (X) components

B: Selection of immune phenotype variables

Figure 3.2 – Explanation of the immune phenotype variability by the sPLS model
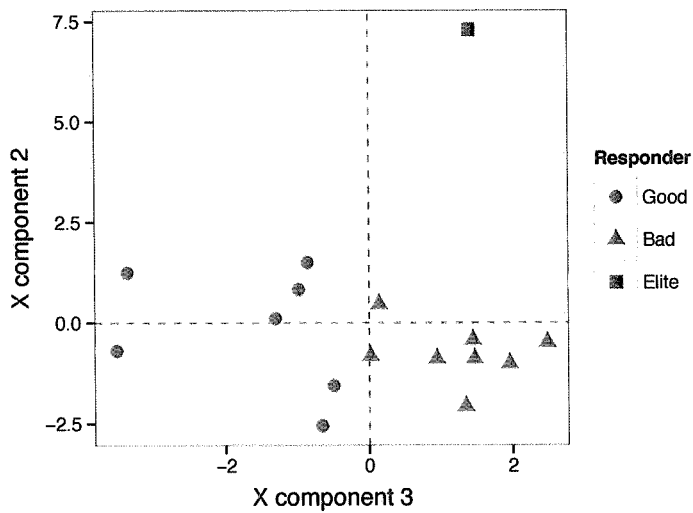
Figure 3.3 – Projection of the patients in the subspace of the 2nd and 3rd sPLS X components

Illumina® probe ID	Gene Module & Symbol	1st comp.	2nd comp.	3rd comp.	4th comp.
ILMN_1663422	M3.2:Inflammation RGL4	-0.02	0.00	0.00	0.00
ILMN_1707312	M3.2:Inflammation NFIL3	-0.19	0.00	0.00	0.00
ILMN_1714592	M3.2:Inflammation CDA	-0.12	0.00	0.00	0.00
ILMN_1911677	M4.1:T cell	0.54	0.00	0.00	0.00
ILMN_1740864	M4.13:Inflammation TREML2	-0.17	0.00	0.00	0.00
ILMN_1661461	M4.13:Inflammation LOC283547	-0.02	0.00	0.00	0.00
ILMN_1714444	M4.15:T cells KLF12	0.30	0.00	0.00	0.00
ILMN_1794588	M4.15:T cells DYRK2	0.14	0.00	0.00	0.00
ILMN_1801216	M4.2:Inflammation S100P	-0.00	0.00	0.00	0.00
ILMN_1795428	M5.11:Undetermined WDR59	0.02	0.00	0.00	0.00
ILMN_1724341	M5.11:Undetermined CXorf45	0.03	0.00	0.00	0.00
ILMN_1747052	M5.4:Undetermined ITGA4	0.12	0.00	0.00	0.00
ILMN_1736757	M5.5:Undetermined GNPTAB	0.02	0.00	0.00	0.00
ILMN_1749892	M5.7:Inflammation EGLN1	-0.04	0.00	0.00	0.00
ILMN_1752932	M5.7:Inflammation MPZL2	-0.03	0.00	0.00	0.00
ILMN_1653432	M5.8:Undetermined HNRPDL	0.38	0.00	0.00	0.00
ILMN_1792173	M5.8:Undetermined TUBGCP4	0.03	0.00	0.00	0.00
ILMN_1801043	M7.1:Inflammation GSN	-0.23	0.00	0.00	0.00
ILMN_1667476	M7.1:Inflammation LTBR	-0.39	0.00	0.00	0.00
ILMN_1746171	M7.1:Inflammation H2AFY	-0.25	0.00	0.00	0.00
ILMN_1695763	M7.11:Undetermined PDIA5	-0.21	0.00	0.00	0.00
ILMN_1740486	M7.14:Undetermined POLR2J4	0.12	0.00	0.00	0.00
ILMN_1743570	M7.16:Undetermined CEACAM3	-0.04	0.00	0.00	0.00
ILMN_1654516	M7.27:Undetermined TMEM120A	-0.15	0.00	0.00	0.00
ILMN_1815306	M3.1:Erythrocytes AP2A1	0.00	-0.08	0.00	0.00
ILMN_1670570	M3.1:Erythrocytes MXI1	0.00	-0.22	0.00	0.00
ILMN_1781001	M4.2:Inflammation SOCS3	0.00	0.01	0.00	0.00
ILMN_1815303	M5.13:Undetermined LOC642197	0.00	0.17	0.00	0.00
ILMN_1652787	M5.2:Undetermined PIK3AP1	0.00	0.06	0.00	0.00
ILMN_1653143	M5.4:Undetermined ECD	0.00	-0.42	0.00	0.00
ILMN_1741881	M5.7:Inflammation C9orf72	0.00	0.46	0.00	0.00
ILMN_1696065	M6.10:Undetermined SDF4	0.00	-0.13	0.00	0.00
ILMN_1785191	M6.9:Undetermined TMEM14A	0.00	0.59	0.00	0.00
ILMN_1807633	M7.11:Undetermined HRSP12	0.00	0.18	0.00	0.00
ILMN_1667171	M7.2:Undetermined LOC651881	0.00	-0.04	0.00	0.00
ILMN_1762883	M7.4:Undetermined ECE2	0.00	0.33	0.00	0.00
ILMN_1784785	M7.5:Undetermined COPS7B	0.00	-0.14	0.00	0.00
ILMN_1685005	M4.13:Inflammation TNFRSF1A	0.00	0.00	0.07	0.00
ILMN_1757730	M4.7:Cell Cycle TTC27	0.00	0.00	-0.47	0.00
ILMN_1708782	M5.1:Inflammation MFAP3	0.00	0.00	0.23	0.00
ILMN_1812688	M5.13:Undetermined C2orf18	0.00	0.00	-0.56	0.00
ILMN_1712423	M5.3:Undetermined SKIP	0.00	0.00	-0.62	0.00
ILMN_1765326	M6.10:Undetermined DGKD	0.00	0.00	0.00	0.00
ILMN_1733441	M7.14:Undetermined POGZ	0.00	0.00	-0.13	0.00
ILMN_1798612	M7.16:Undetermined SNX20	0.00	0.00	-0.03	0.00
ILMN_1717337	M5.7:Inflammation MARCH7	0.00	0.00	0.00	1.00

Table 3.1 – Loadings of the selected genes on any of the 4 X sPLS components
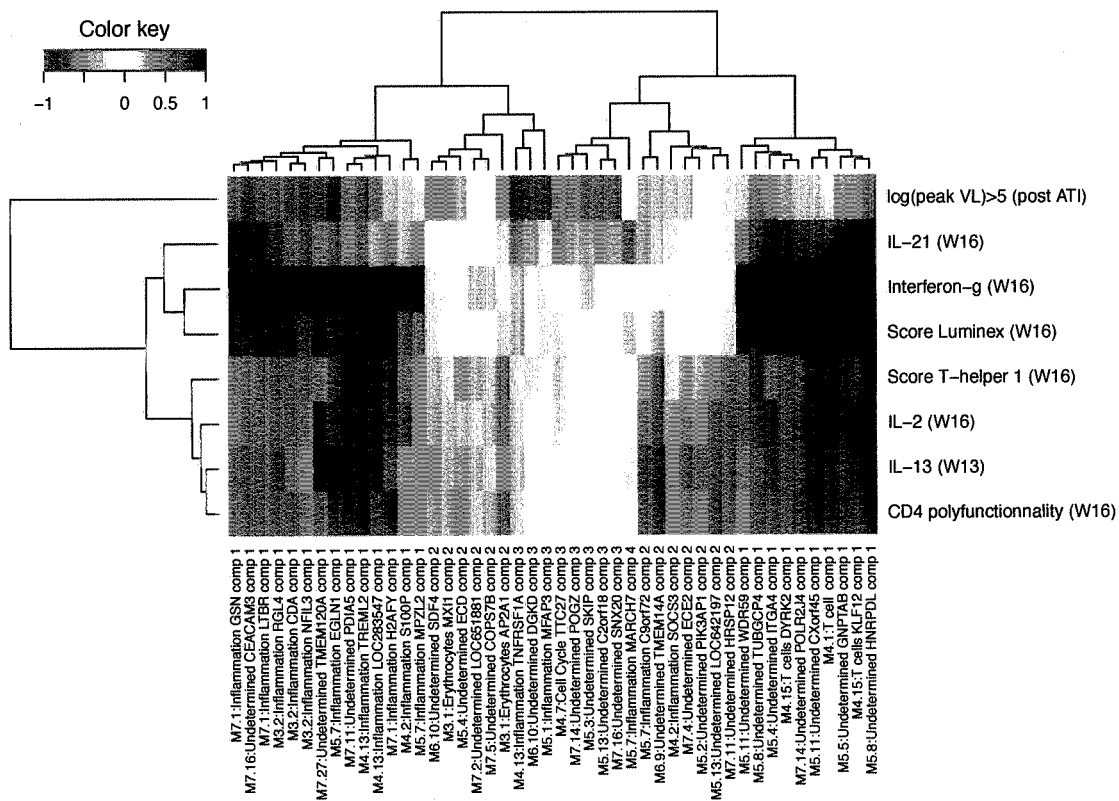
Figure 3.4 – Correlations between variables and genes selected by one of the first four components
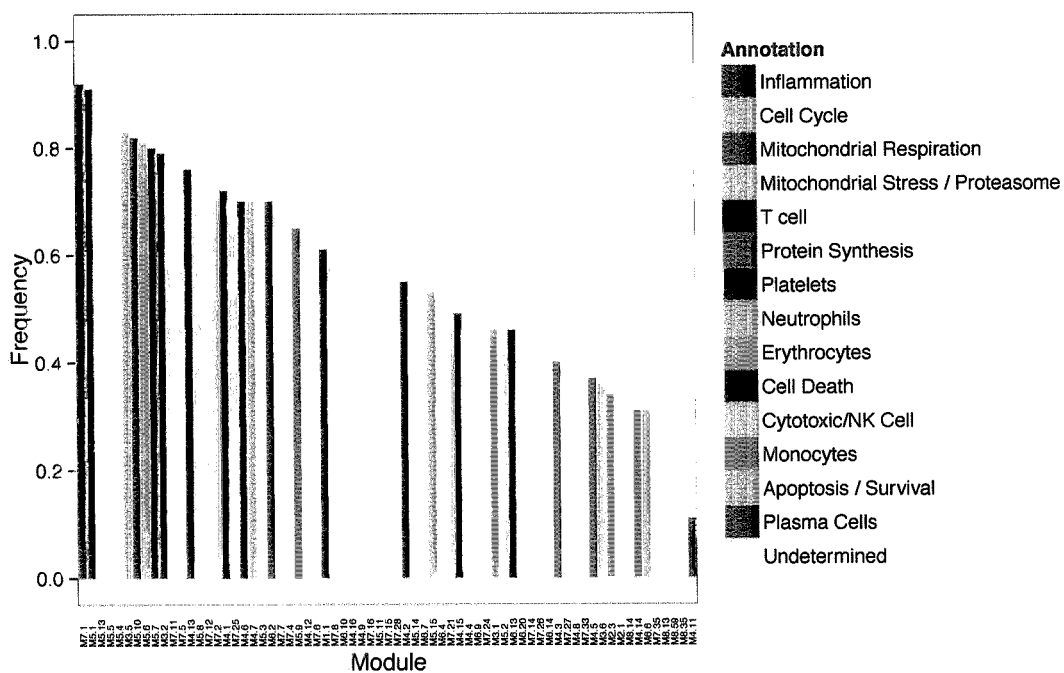
Figure 3.5 – Modules selection Robustness on 100 bootstrap samples

All in all it seems that the reproducibility of the variable selection at the gene level is quite low. However, reproducibility at the Module level is relatively good, with the stronger signal (Inflammation, T-cell activation, ...) being very consistent (Figure 3.5 and Table 3.2). Once again, many selected Modules are not annotated. In addition, VSURF and bootstrap results were much in agreement.

Conclusions

Finally changes in gene expression in whole blood were consistently associated with results from in vitro T cell assays. The integration of gene expression patterns and functional assays identified signature inversely correlated with the peak of viral load following ATI. Indeed patients with lower expression of inflammatory pathways during vaccination developed a stronger vaccine elicited immune response. This analysis may help to better define the endpoint of next vaccine trials and to identify patients who likely benefit from the vaccination.

3.2 Systems analysis of sex differences in the response to influenza vaccination

3.2.1 Introduction to sex variability in immunity

The variability in the biology of human populations rises significant challenges in understanding various disease outcomes and developing successful therapeutics. The sources of this variation are likely the consequence of genetics, epigenetics, and the history of antigenic exposure [Jirtle and Skinner, 2007; Knight, 2013]. As therapies targeting immune function are developed to improve clinical outcomes in many situations such as cancer,

Module	Freq	Annotation	Module	Freq	Annotation
M7.1	92%	Inflammation	M7.28	56%	Undetermined
M5.1	91%	Inflammation	M4.2	55%	Inflammation
M5.13	91%	Undetermined	M5.14	55%	Undetermined
M5.5	91%	Undetermined	M6.7	55%	Undetermined
M5.4	84%	Undetermined	M5.15	53%	Neutrophils
M3.5	83%	Cell Cycle	M6.4	52%	Undetermined
M5.10	82%	Mitochondrial Respiration	M7.21	50%	Undetermined
M5.6	81%	Mitochondrial Stress / Proteasome	M4.15	49%	T cell
M5.7	80%	Inflammation	M4.4	49%	Undetermined
M3.2	79%	Inflammation	M6.9	48%	Undetermined
M7.11	79%	Undetermined	M7.24	48%	Undetermined
M7.5	78%	Undetermined	M3.1	46%	Erythrocytes
M4.13	76%	Inflammation	M5.2	46%	Undetermined
M5.8	76%	Undetermined	M6.13	46%	Cell Death
M7.12	76%	Undetermined	M6.20	46%	Undetermined
M7.2	73%	Undetermined	M7.14	46%	Undetermined
M4.1	72%	T cell	M7.26	46%	Undetermined
M7.25	72%	Undetermined	M6.14	42%	Undetermined
M4.6	70%	Inflammation	M4.3	40%	Protein Synthesis
M4.7	70%	Cell Cycle	M7.27	40%	Undetermined
M5.3	70%	Undetermined	M4.8	38%	Undetermined
M6.2	70%	Mitochondrial Respiration	M7.33	38%	Undetermined
M7.7	70%	Undetermined	M4.5	37%	Protein Synthesis
M7.4	68%	Undetermined	M3.6	36%	Cytotoxic/NK Cell
M5.9	65%	Protein Synthesis	M2.3	34%	Erythrocytes
M4.12	64%	Undetermined	M2.1	33%	Undetermined
M7.6	64%	Undetermined	M8.14	32%	Undetermined
M1.1	61%	Platelets	M4.14	31%	Monocytes
M7.8	61%	Undetermined	M6.6	31%	Apoptosis / Survival
M6.10	60%	Undetermined	M7.35	26%	Undetermined
M4.16	58%	Undetermined	M8.13	21%	Undetermined
M4.9	58%	Undetermined	M8.59	21%	Undetermined
M7.16	58%	Undetermined	M8.35	17%	Undetermined
M5.11	56%	Undetermined	M4.11	11%	Plasma Cells
M7.15	56%	Undetermined			

Table 3.2 – Modules selection consistency on 100 bootstrap samples

infections, autoimmune diseases and transplantation, identifying the sources of immunological variation and finding biomarkers associated with immune health are crucial for their success [Davis, 2008]. An important source of immunological variation is known to be the sex of the individual. Males experience a greater severity and prevalence of bacterial, viral, fungal, and parasitic infections than females, who also exhibit a more robust response to antigenic challenges such as infection and vaccination [Klein, 2000; Klein and Poland, 2013]. This stronger immune response in females could also explain why they more frequently develop immune-mediated pathologies during influenza infection [Robinson et al., 2011]. Furthermore, females are at a higher risk for developing autoimmune diseases. In this later context, it is interesting to note that a recent study showed that females had, on average, 1.7 times the frequency of self-specific T cells as males [Su et al., 2013]. Despite the fact that initial observations relating the sex of the individual with the immune response were made many years ago [Grossman, 1985], little is known about the mechanisms underlying these differences. Some sex-specific variations in the immune response can be directly attributed to sex hormones [Sakiani et al., 2013]. In humans, sex steroids can bind to intracellular receptors located in immune cells such as monocytes, B cells, and T cells and activate hormone-responsive genes, suggesting that they can directly affect sex-related differences in both innate and adaptive immune responses [Pennell et al., 2012]. Whereas estrogens are associated with inflammation and can stimulate proliferation and differentiation of lymphocytes and monocytes, androgens suppress the activity of immune cells by increasing the synthesis of anti-inflammatory cytokines [Olsen and Kovacs, 1996; Liva and Voskuhl, 2001]. To date, no clear associations have been found between biological and clinical differences in the immune response between males and females in humans. In one study, results from public gene expression data [Gaucher et al., 2008] showed that many of the genes induced by a yellow fever vaccine were preferentially activated in females [Klein et al., 2010]. However, whether these differences correlate with poor antibody outcomes remains to be determined.

3.2.2 Serological response to trivalent inactivated seasonal influenza vaccine

To study the differences in males' versus females' immune systems, we used data from a vaccination and systems immunology study conducted on 91 individuals (37 males and 54 females) of different ages (20 to 30 and 60 to 89 years old) that was recently reported [Furman et al., 2013]. We studied a variety of immune parameters from peripheral blood before vaccination, including cytokines, chemokines, and growth factors in serum, frequencies of diverse blood cell subsets, phosphorylation levels of signal transducer and

activator of transcription (STAT) proteins in multiple cells stimulated with a variety of cytokines or unstimulated (96 conditions in total), and whole-blood gene expression. The gene expression data were reduced to 109 gene modules by cluster analysis and assignment of a set of transcription factors (regulatory program) to each gene module as described [Furman et al., 2013] (Appendix D.1 page 141). Four individuals were removed from the analysis: two outliers and two with incomplete dataset.

In this study, we sought to determine whether we could identify biomarkers from peripheral blood that could explain the sex-related differences in the serological response to the trivalent inactivated seasonal influenza vaccine (TIV) in both young (20 to 30 years old) and older cohorts (60 to 89 years old). To determine the magnitude of the antibody response to TIV, virus microneutralization assays were performed for each of the three strains contained in the TIV (H1N1, H3N2, B). The seroconversion rate (percent of individuals with a fourfold or greater change in their post- versus pre-vaccination antibody microneutralization titer) was computed for each strain and for each group of individuals (Figure 3.6). Young and older females had higher neutralizing antibodies than age-matched males (Figure 3.6), as previously reported [Cook, 2008]. The largest differences between males and females were observed for the H3N2 strain (Figure 3.6 and Table 3.3). Females also showed higher expression of inflammatory markers, however, none of these specific sex-related differences correlated with the observed disparities in the antibody response to TIV.

In the rest of the analysis, the focus was only on the antibody response to the H3N2 strain.

3.2.3 Interaction analysis and modeling of antibody response to the H3N2 strain

The goal of this analysis was to identify potential gene module expressions that were related to the sex effect on the antibody response to TIV for the H3N2 strain. The response was modeled as a binary variable (fold increase of post- versus pre-vaccination antibody microneutralization titer ≥ 4). A logistic regression [McCullagh and Nelder, 1989] was conducted for the estimation of the regression coefficients and odds ratios in response to vaccination.

First and foremost, potential confounders of the sex effect on the response were considered. A variable was labelled as a confounder of the sex effect on the antibody response if it modified the estimate of the sex effect on the response by more than 20%. A forward strategy was performed starting with a basic model including the sex covariate only. Two potential confounders were thus identified: the gene module 42 which is linked to the
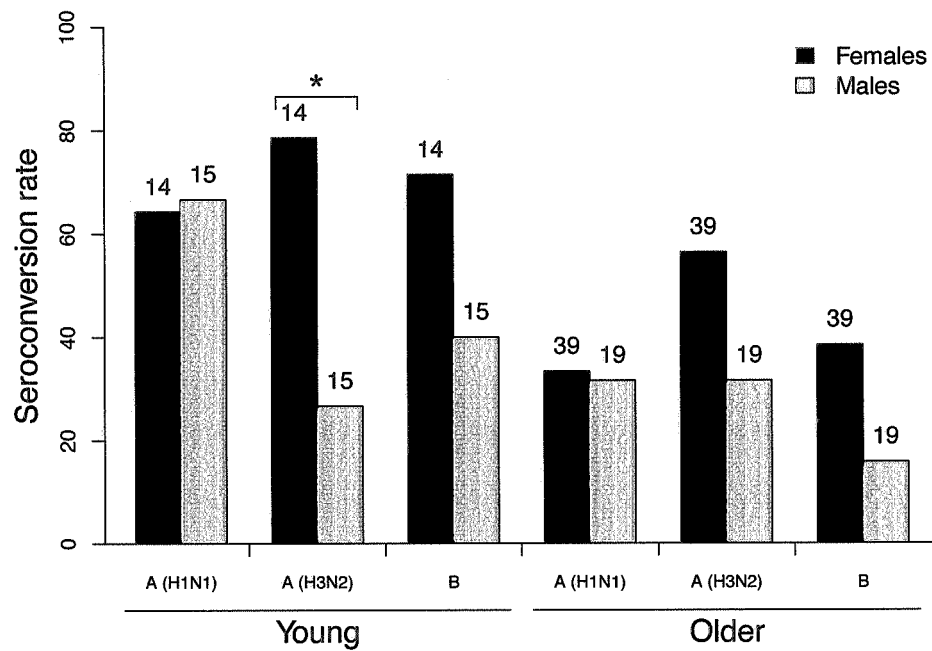
⋆: chi-square test p-value < 0.05

Note: numbers above the bars are the total number of individuals in each category.

Figure 3.6 – Seroconversion rate after TIV in all three stains stratified on sex for young and old individuals

Strain	Variable	Effect size	p-value
H1N1	(intercept)	-0.272	0.234
	age	-0.690	0.004⋆
	male	-0.011	0.962
H3N2	(intercept)	-0.038	0.962
	age	-0.190	0.421
	male	-0.716	0.003⋆
B	(intercept)	-0.502	0.033⋆
	age	-0.583	0.018⋆
	male	-0.594	0.020⋆

⋆: significant at a 5% level for type I error

Note: The response was modeled as a binary variable (fold increase of post- versus pre-vaccination antibody microneutralization titer ≥ 4) in a logistic regression model [McCullagh and Nelder, 1989]. For the sex variable *male*, women are the reference.

Table 3.3 – Age and sex effects on antibody titer responses to TIV

encoding of ribosomal proteins and the acute-phase inflammatory marker CRP (Figure 3.7). First gene module 42 modified the sex effect by more than 20% in the following model:

$$\text{logit} (p(y_i = 1)) = \mu + \beta_s male_i + \beta var_i \tag{6}$$

where $y_i = \mathbb{1}_{\{AntibodyTiter(prevac)/AntibodyTiter(postvac) \geq 4\}}$, μ is the intercept, β_s is the effect of being male (female are the reference) on the logit of the probability of responding to TIV for the H3N2 strain, $male_i$ is 1 if individual i is male and 0 if individual i is female, β the effect of the potential confounder variable, and var_i the potential confounder variable value for individual i. The model (6) was tried out for all immune variables in place of var. Gene module 42 was the variable with the largest modification of the sex effect. Then CRP modified the sex effect by more than 20% in the following model:

$$\text{logit} (p(y_i = 1)) = \mu + \beta_s male_i + \beta_{m42} m42_i + \beta var_i \tag{7}$$

Then no more variables modified the sex effect on the response once adjusting on both CRP and gene module 42 using again the same strategy (Figure 3.7) and the final model adjusted on potential confounders was:

$$\text{logit} (p(y_i = 1)) = \mu + \beta_s male_i + \beta_c CRP_i + \beta_{m42} m42_i \tag{8}$$

Second, to identify possible gene module candidates that explain the differences observed in responsiveness, significant marginal interactions between gene modules and the sex effect on the response were tested. To do so, the **Interact** strategy from Simon and Tibshirani [2012] was used in order to investigate all possible interactions in the binary response while controlling the FDR using permutation methods. Testing for interactions in high-dimensional settings is difficult and can lead to several issues. In particular, applying multiple testing correction to multiple fitted bivariate logistic regression models is problematic [Bůžková et al., 2011]. Indeed, traditional FDR controlling procedures (Appendix A page 121) can fail to control the FDR in the logistic regression settings [Simon and Tibshirani, 2012]. In addition, it is difficult to derive a permutation strategy that would test only the interaction effect, and not both main and interaction effects at once in logistic regression models under reasonable assumption on the independence of both variables at play [Bůžková et al., 2011]. Simon and Tibshirani [2012] proposes a backward strategy for testing, for each pair of variables, if their correlation is the same regardless of the class output (in our case). Their strategy allows to derive a permutation method that controls the FDR correctly [Simon and Tibshirani, 2012] and thus identify any significant marginal interactions. Setting the significance threshold at an FDR of $< 10\%$ identified only one gene module interacting with the sex on the antibody response against the H3N2 strain: gene module 52.
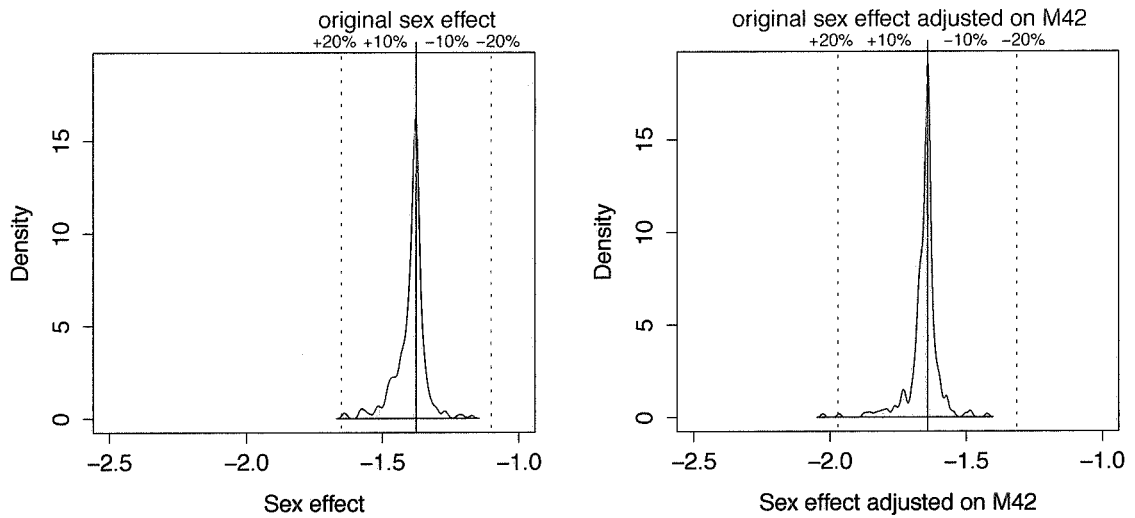
A: Density of sex effects on antibody response: $\text{logit}\,(p(y_i = 1)) = \mu + \beta_s male_i + \beta var_i$

B: Density of sex effects on antibody response adjusted gene module 42 : $\text{logit}\,(p(y_i = 1)) = \mu + \beta_s male_i + \beta_{m42} m42_i + \beta var_i$
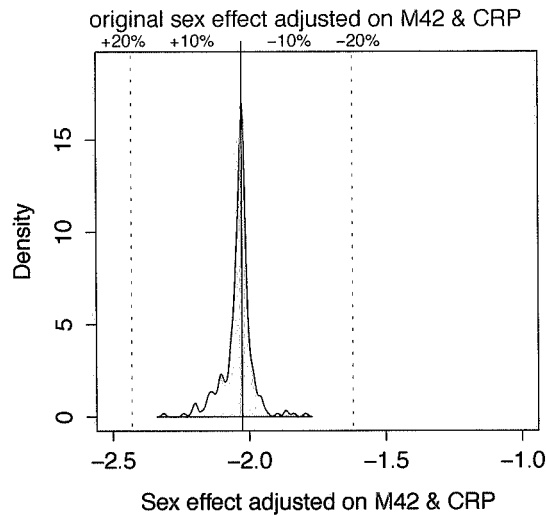
C: Density of sex effects on antibody response adjusted on gene module 42 and CRP: $\text{logit}\,(p(y_i = 1)) = \mu + \beta_s male_i + \beta_{m42} m42_i + \beta_c crp_i + \beta var_i$

Note: *var* designate the potential confounders tested (one variable at a time). Negative values (x axis) indicate higher vaccine response in females (the reference).

Figure 3.7 – Densities of size effects of the sex variable on the antibody response to H3N2 strain after TIV when ajusting on various potential confounders

The following model was thus estimated (model (9)), which included the variables of sex and gene module 052 and their interaction term, adjusted on the potential confounding covariables gene module 42 and CRP:

$$\text{logit}\,(p(y_i = 1)) = \mu + \beta_s male_i + \beta_c CRP_i + \beta_{m42} m42_i \tag{9}$$
$$+ \beta m52_i + \beta_{s:v} male_i : m52_i$$

The resulting odds ratio (OR) estimate for vaccine response based on the expression of gene module 52 in model (9) was 0.39 [confidence interval (CI), 0.18-0.84] for males and 2.25 (CI, 1.08-4.67) for females (Figure 3.8A). This indicates that the probability of being a high responder to TIV for the H3N2 strain significantly decreases with an elevated expression of module 052 in males and with decreased expression of module 52 in females. To determine the extent to which module 052 and its interaction with sex contribute to the classification model, we computed a cross-validated (leave-pair-out [Airola et al., 2011]) area under the curve (AUC_{cv}) for model (8) and model (9). The AUC_{cv} was 0.712 for model (8), and 0.761 for model (9). Furthermore, direct comparison of the two models by a likelihood ratio test showed that model (9) is significantly better (p-value of 0.0019) than model (8). These results suggest that the observed sex differences in the neutralizing antibody responses to vaccination could be mediated by the expression of genes involved in lipid metabolism (as a significant part genes participating in the gene module 52 are involved in lipid metabolism).

Our results showing that augmented expression of gene module 52 correlated with weaker TIV responsiveness in males but not in females suggested that sex hormones could be involved in expression of this gene module. Indeed, results from chemical-gene interaction analysis (http://ctdbase.org) (24) show that expression of a significant fraction of genes in module 052 can be modulated by testosterone (hypergeometric test p-value < 0.005). Free (unbound, bioactive form) testosterone was measured in the sera from the individuals in our study with the hypothesis that, in males, the observed effect of module 52 on vaccine response was dependent on the circulating levels of testosterone. Male subjects were stratified into testosterone high (Thi) or low (Tlo) if they were respectively above or below the median for all of the male subjects (in our sample testosterone median is 4.06 pg/mL and testosterone ranges from 0.58 pg/mL to 24.78 pg/mL). A final logistic regression model (model 10) for antibody response to H3N2 strain was estimated, in which the sex variable was replaced by a three category variable: individuals were either female (the reference), Thi male (n = 17) or Tlo male (n = 17). The median testosterone level in Thi subjects was 9.55 pg/mL (ranging from 4.25 pg/mL to 24.78 pg/mL), and 2.34 pg/mL (ranging from 0.58 pg/mL to 3.89 pg/mL) in Tlo subjects. The median age for Tlo and Thi males was 77 and 24 years, respectively. Thus, model 3 included the interaction
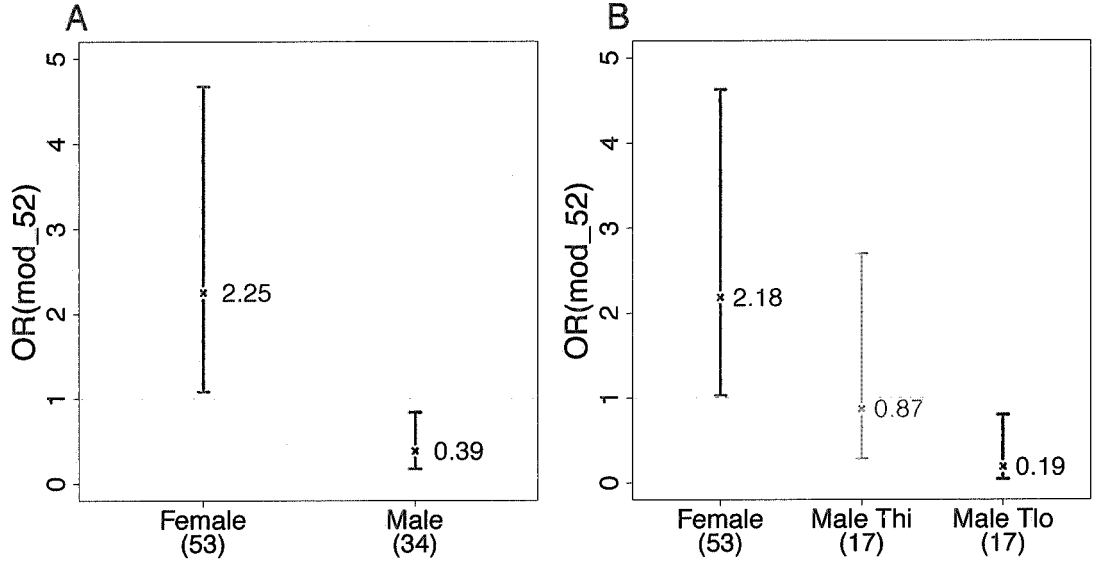
Note: Interaction analysis was conducted for sex and gene expression modules on the serological (antibody microneutralization) responses to TIV (seroconversion to the H3N2 strain). A significant interaction was found between the variables sex and gene module 052. (A) Odds ratio for vaccine response given module 052 in females (red line) and males (blue line). (B) No significant interaction between sex and module 052 was observed for males with low levels of testosterone [Tlo (n = 17), brown line], although a significantly negative effect of module 052 was observed for males with high levels of testosterone [Thi (n = 17), blue line] (adjusted for confounders including age)

Figure 3.8 – Odds ratios of gene module 52 for antibody responses to H3N2 strain after TIV based on gender and Testosterone level

terms gene module 52 *times* Thi and gene module 52 × Tlo, and was also adjusted for age, because of the effect of aging on testosterone levels:

$$\text{logit}\,(p(y_i = 1)) = \mu + \beta_c crp_i + \beta_{m42} m42_i + \beta_{m52} m52_i + \beta_a age_i \tag{10}$$
$$+ \beta_{Tlo} maleTlo_i + \beta_{Tlo:m52} maleTlo_i : m52_i$$
$$+ \beta_{Thi} maleThi_i + \beta_{Thi:m52} maleThi_i : m52_i$$

Strikingly, the interaction between testosterone levels and gene module 52 was significant only for the Thi group (Wald test p-value < 0.005 – see Figure 3.8B), and not for Tlo males (Wald test p-value = 0.18), and the corresponding OR estimates for vaccine response, according to module 052, were 0.87 (CI, 0.28-2.69) for Tlo and 0.19 (CI, 0.04-0.80) for Thi males. We also tested testosterone levels as a continuous measure in the following model (also adjusted by sex and age):

$$\text{logit}\,(p(y_i = 1)) = \mu + \beta_c crp_i + \beta_{m42} m42_i + \beta_{m52} m52_i + \beta_a age_i \tag{11}$$
$$+ \beta_s male + \beta_T testo_i + \beta_{T:m52} testo_i : m52_i$$

Consistent with model 10, the interaction of gene module 52 and continuous testosterone levels was significant (Wald test p-value = 0.012 – see the significantly decreasing trend of the OR of gene module 52 as testosterone level increase in Figure 3.9). This indicates that gene module 52 has a significant effect on antibody response after TIV in males with high levels of testosterone but not in those with lower levels. Models 10 and 11 were not nested therefore likelihood ratio test is not available to compare them. Both AIC and BIC criteria favor model 11 (respectively 102.9 and 120.0) over model 10 (respectively 106.3 and 128.5). However, model 10 is more interpretable as including testosterone level of women in the estimation is debatable.
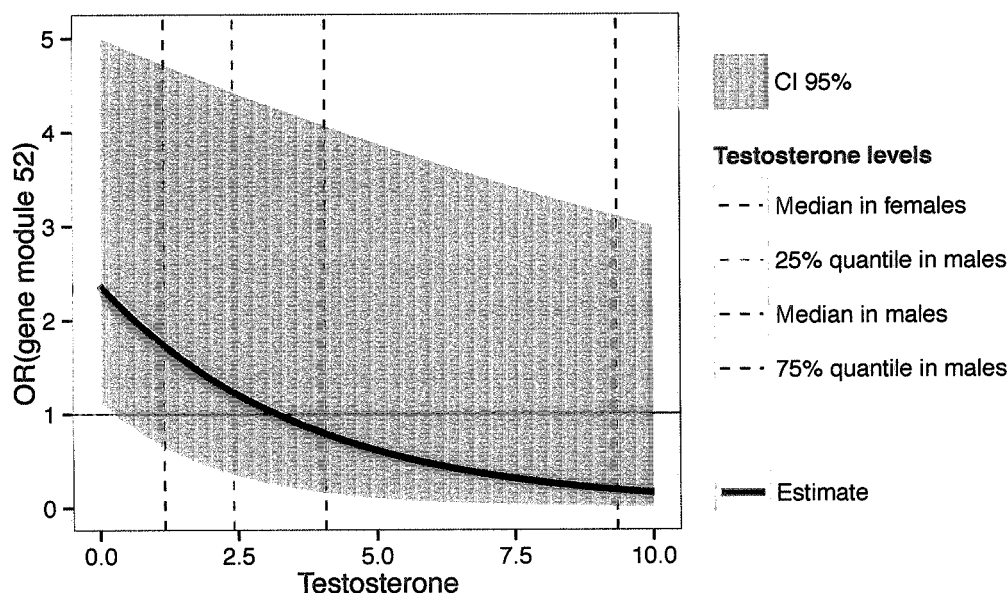
Note: Odds ratio for vaccine response based on expression of module 052 and testosterone levels. Logistic regression analysis was conducted on the antibody-neutralizing activity based on expression of genes in module 052 and the testosterone levels as a continuous measurement. The estimated odds ratio (OR) for the antibody-neutralizing response is shown (black continuous line). Red, light blue, blue, and dark blue dashed lines indicate the median testosterone levels in females and first, second, and third quartiles of testosterone levels for males. CI, confidence interval.

Figure 3.9 – Odds ratios of gene module 52 for vaccine responses based on expression of module 52

Together, these results show that in males with higher levels of testosterone and elevated expression of gene module 52 that participates in lipid metabolism, the antibody response to vaccination is severely down-regulated, whereas in those with low levels of testosterone, or in females, the contribution of gene module 52 is not detrimental and the responses to the vaccine remain intact.

Conclusions

There are marked differences between the sexes in their immune response to infections and vaccination, with females often having significantly higher responses. However, the mechanisms underlying these differences are largely not understood. Using a systems immunology approach, we have identified a cluster of genes involved in lipid metabolism and likely modulated by testosterone that correlates with the higher antibody-neutralizing response to influenza vaccination observed in females. Moreover, males with the highest testosterone levels and expression of related gene signatures exhibited the lowest antibody responses to influenza vaccination. This study generates a number of hypotheses on the sex differences observed in the human immune system and their relationship to mechanisms involved in the antibody response to vaccination.

4 Dirichlet process mixture of skew t-distributions for modeling flow cytometry data

Abstract:

Flow cytometry is a high-throughput technology used to quantify multiple surface and intracellular markers at the level of a single cell. Improvements of this technology lead today to the ability of describing millions of individual cells from a blood sample using multiple markers. This allows to identify cell sub-types, and to count the number of cells of each sub-type. But it also results in high-dimensional datasets, whose manual analysis is highly time-consuming and poorly reproducible. Several methods have been developed to perform automated recognition of cell populations from flow cytometry data. Most of them are suited for the analysis of a single sample from one patient. In clinical trials, repeated measurements with several samples by patient and by time points are actually available. We propose to use a Bayesian nonparametric approach with Dirichlet process mixture (DPM) of skew t-distributions to perform model based clustering of such data. DPM models enable the number of cell populations to be estimated from the data avoiding any model selection. The use of skew t-ditributions provides robustness to outliers and suits best the usually non elliptical shape of cell population distributions. In the case of repeated measurements, we propose a sequential strategy relying on a parametric approximation of the posterior. We apply this methodology to simulated data and to two experimental benchmark datasets.

Key Words: Automated gating; Bayesian; Dirichlet process; Flow cytometry; Mixture model; Nonparametrics Bayesian; Skew t-distribution.

Contents

Valorisation: section 4.4 is mainly part of an article that is in preparation for submission for publication in a peer reviewed scientific journal.

4.1 Introduction to flow cytometry data

Flow cytometry is a high-throughput technology used to quantify multiple surface and intracellular markers at the level of single cell. More specifically, cells are stained with multiple fluorescently-conjugated monoclonal antibodies directed to cell surface receptors (such as CD4) or intracellular markers (such as the interleukine-2 cytokine) to determine the type of cell, their differentiation and their functionality. Figure 4.1 shows a simplistic representation of this idea. A flow cytometer (Figure 4.2) is then used to measured the fluorescent intensity of the stained cells one by one. With the improvement of this technology, leading currently to the use of up to 18 markers at the same time (using 18 colors), multi-parametric description of millions of individual cells can be generated.

4.2 A brief introduction to the Bayesian framework

The Bayesian approach to probability theory dates back to the 18th century and the posthumous publication of Bayes article in 1764 (even though the discovery of Bayes's rule is subject to historical controversies [Stigler, 1983]). The Bayesian approach is often opposed to the frequentist one. The frequentist approach assumes that model parameters are deterministic, having a fixed (unknown) value, whereas in the Bayesian framework, parameters themselves are considered as random variables, with associated probabilities.

Let's consider the following model where the observations y and the parameters θ have a joint probability density function $p(\theta, y)$. This joint density can be decomposed as the product of the prior density function (i.e. the probability distribution representing the prior belief or knowledge on the parameters) $p(\theta)$ and the likelihood of the data $p(y|\theta)$:

$$p(\theta, y) = p(\theta)p(y|\theta)$$

Using the same decomposition but with the marginal distribution of the observations $p(y)$, one gets:

$$p(\theta, y) = p(y)p(\theta|y)$$

Then Bayes' Theorem can be easily derived, giving the posterior density of θ:

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} = \frac{p(\theta)p(y|\theta)}{\int p(\theta)p(y|\theta))}$$

The purpose of Bayesian inference is to estimate this posterior density of the parameters, usually focusing only on the part dependent of θ:

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$
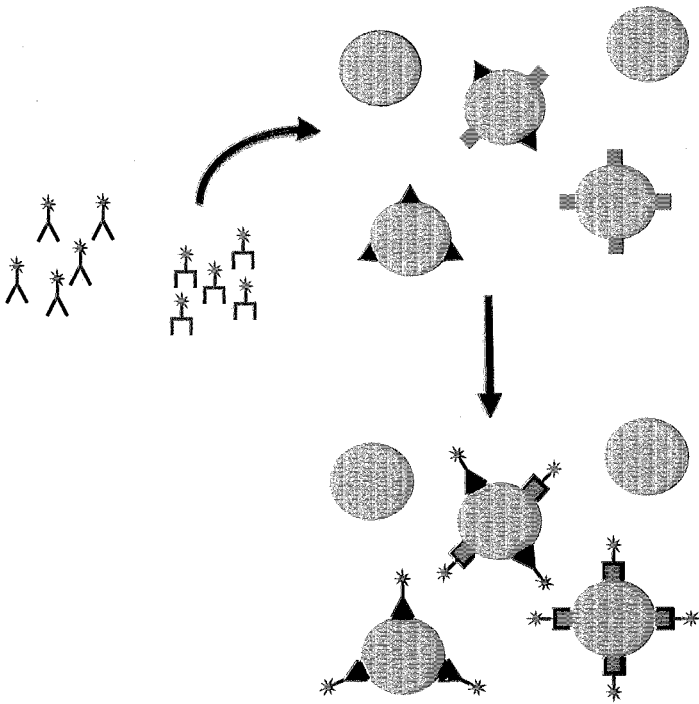
Figure 4.1 – Simplistic representation of the idea of cellular markers staining
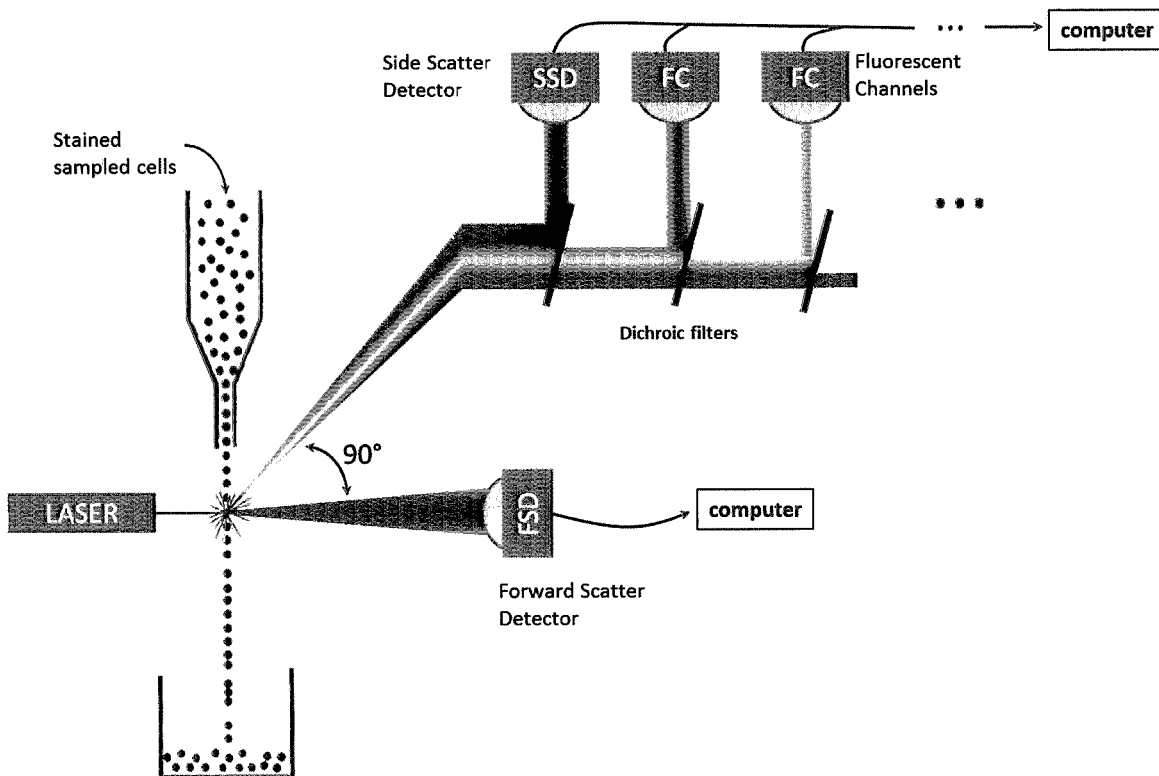
Figure 4.2 – Simplistic representation of a flow cytometer

Various techniques can be used to estimate the posterior distribution of the parameters [Gelman et al., 2013]. Posterior simulations through Markov Chain Monte Carlo (MCMC) methods [Geman and Geman, 1984; Gilks et al., 1996] have become extremely useful thanks to the increased computing capabilities of modern computers.

4.3 Dirichlet process mixture models

4.3.1 Mixture Models

Let's consider C observations in d dimensions: $y_c \in \mathbb{R}^d$, $c = 1, \ldots, C$ (typically corresponding to the vector of fluorescence intensities measured for the cell c in the case of flow cytometry data). We assume that those data are independent and identically distributed (i.i.d.) from some unknown distribution F:

$$y_c | G \overset{iid}{\sim} F \text{ for } c = 1 \ldots, C \tag{12}$$

where F is a mixture of distributions:

$$F(y) = \int_{\Theta} f_{\theta}(y) G(d\theta) \tag{13}$$

where $f_{\theta}(y)$ is a known probability density function, parameterized by $\theta \in \Theta$, a set of parameters, and defining the shape of a cluster. G is the unknown mixing distribution, which carries the weights and locations of the mixture components. In a parametric approach, $G = \sum_{k=1}^{K} \pi_k \delta_{\theta_k}$ where π_k is the weight of the k^{eth} mixture component and θ_k its respective parameters. Maximum likelihood or Bayesian estimates of F can be derived for such models [Biernacki et al., 2000].

4.3.2 Dirichlet process mixture models

Dirichlet Process definition and basic properties

In a nonparametric perspective (where the number of clusters is unknown) G is written as a infinite sum of atoms: $G = \sum_{k=1}^{+\infty} \pi_k \delta_{\theta_k}$. Let's assume that the random mixing distribution G is drawn from a Dirichlet process [Ferguson, 1973]:

$$G \sim \text{DP}(\alpha, G_0) \tag{14}$$

where $\text{DP}(\alpha, G_0)$ denotes the Dirichlet process of scale parameter $\alpha > 0$ and base probability distribution G_0.

Definition. A Dirichlet Process (DP) is a probability distribution over the space of probability measures whose marginal distributions are Dirichlet distributed: If $G \sim \mathrm{DP}(\alpha, G_0)$, $\forall$ partition $A_1, \ldots, A_k$ of a measurable space $\mathcal{T}$,

$$(G(A_1), \ldots, G(A_k)) \sim \mathrm{Dir}\,(\alpha G_0(A_1), \ldots, \alpha G_0(A_k))$$

The DP constitutes a very appealing prior for unsupervised clustering models, as it generates infinitely large atomic discrete distributions. Indeed, a draw $G \sim \mathrm{DP}(\alpha, G_0)$ is almost surely discrete and takes the following form [Sethuraman, 1994]:

$$G = \sum_{k=1}^{+\infty} \pi_k \delta_{\theta_k} \tag{15}$$

where the θ_k are i.i.d. from the base distribution G_0 and independent of the weights, $\boldsymbol{\pi} = (\pi_k)_{k=1,2,\ldots}$, which are drawn from a so-called "stick-breaking" distribution:

$$\pi_1 = \beta_1 \quad \text{and} \ \forall k > 1, \ \pi_k = \beta_k \prod_{j=1}^{k-1}(1 - \beta_j)$$

with $\beta_k \overset{iid}{\sim} \mathrm{Beta}(1, \alpha)$ for $k = 1, 2, \ldots$. We write simply $\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha)$ after the Griffiths-Engen-McCloskey (GEM) distribution [Pitman, 2006]. The base distribution G_0 tunes the prior information available about the cluster locations. The parameter α tunes the prior distribution on the overall number of clusters K that will be discovered within C data. In particular we have $\mathbb{E}[K|C] = \sum_{c=0}^{C-1} \frac{\alpha}{\alpha+C}$ and $V[K|C] = \sum_{c=0}^{C-1} \frac{\alpha c}{(\alpha+c)^2}$ [Teh, 2010].

The DP carries conjugacy properties from the finite Dirichlet distribution, and we can write the conditional distribution of $\boldsymbol{\theta}_c$ given $\boldsymbol{\theta}_{-c} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{c-1}, \boldsymbol{\theta}_{c+1}, \ldots, \boldsymbol{\theta}_c)$, known as the Blackwell-MacQueen scheme:

$$\boldsymbol{\theta}_c | \boldsymbol{\theta}_{-c} \sim \frac{\alpha}{\alpha + C - 1} G_0 + \frac{1}{\alpha + C - 1} \sum_{m \neq c} \delta_{\boldsymbol{\theta}_m} \tag{16}$$

where δ_{θ_m} denotes the probability distribution with all its mass in θ_m. Note that the values of the drawn θ are repeated, with the probability of an already observed value to be observed again proportional to the number of times it has already been observed. This can be viewed as a "rich gets richer" property. The partition distribution induced by the clustering can be interpreted as a Chinese Restaurant Process (CRP). This is a metaphor of a Chinese restaurant with an infinite number of tables, and where customers enter one by one and sit to any table they want (each table can also sit an infinite number of customers). A new customer entering the restaurant can sit either at a table where previous customers are already sitting, or sit alone at a new table. The most popular tables

are the ones with the most customers already sat. In this metaphor, the customers are the observations and the tables are in fact the clusters. The CRP describes a generative process for the partition induced by a DP.

The model defined by Equations (12), (13) and (14) yields the following hierarchical model known as a Dirichlet process mixture (DPM) model [Antoniak, 1974; Lo, 1984; Escobar and West, 1995; Teh, 2010] with a Gamma hyperprior on the concentration parameter α:

$$\alpha | a, b \sim \text{Gamma}(a, b) \tag{17a}$$

$$\pi \,|\, \alpha \sim \text{GEM}(\alpha) \tag{17b}$$

for $k = 1, 2, \dots$

$$\theta_k \,|\, G_0 \sim G_0 \tag{17c}$$

for $c = 1, 2, \dots, C$

$$\ell_c \,|\, \pi \sim \text{Mult}(\pi) \tag{17d}$$

$$y_c \,|\, \ell_c, \{\theta_k\} \sim f_{\theta_{\ell_c}} \tag{17e}$$

where ℓ_c is an allocation variable indicating to which cluster is associated the observation c. The parameter α has a direct impact on the number of clusters in the posterior (Figure 4.3). In order to truly estimate the number of clusters in the posterior partition from the data, it is important to specify a hyperprior on α.

4.4 Automated gating of flow cytometry data

This section is mainly part of an article that is in preparation for submission for publication in a peer reviewed scientific journal.

Analysis of flow cytometry data is usually performed manually. This results in analyses: i) poorly reproducible [Aghaeepour et al., 2013], ii) expensive (highly time-consuming) and iii) as a result of ii), focused on specific cell populations (i.e. specific combination of markers), possibly missing out on cell populations. Efforts have been made in the recent years to offer automated solutions to tackle these limitations [Aghaeepour et al., 2013], and a lot of different methodological approaches have been proposed to perform automated recognition of cell populations from flow cytometry data. First, clustering methods related to the k-means methods such as L2kmeans [Aghaeepour et al., 2013],
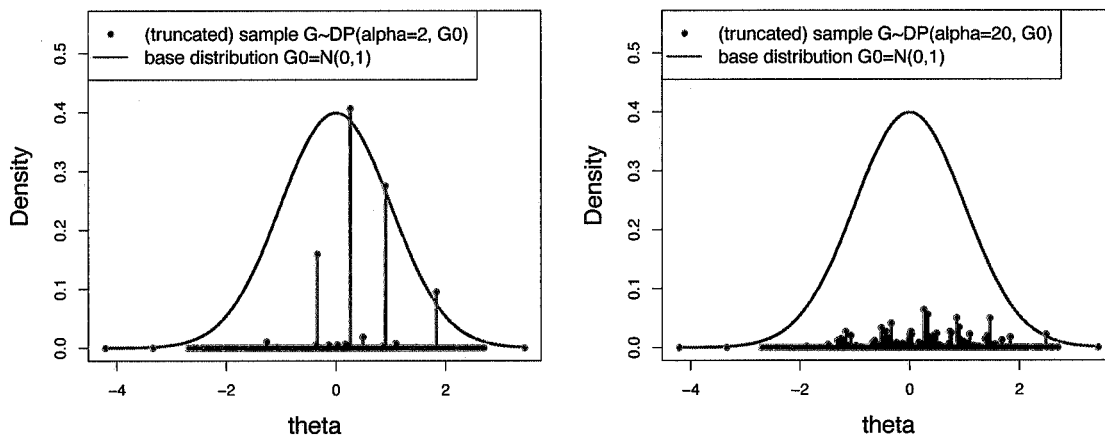
A: Example of a DP sampling with α=2   B: Example of a DP sampling with α=20

Figure 4.3 – Impact of α on the number of locations with a significant weight

flowMeans [Aghaeepour et al., 2011] were proposed. Model based clustering methods relying on finite mixture models such as flowCust/merge [Lo et al., 2008; Finak et al., 2009], FLAME [Pyne et al., 2009], SWIFT [Naim et al., 2014] were also proposed, as well as dimension reduction methods such as MM and MMPCA [Sugár and Sealfon, 2010], Sam-SPECTRAL [Zare et al., 2010], FLOCK [Qian et al., 2010]. All those approaches requires the number of cell populations to be fixed in advance, and resort to various criteria to find the optimal number of cell populations. Finally, CDP [Chan et al., 2008], and more recently [Lin et al., 2013; Cron et al., 2013; Dundar et al., 2014], proposed nonparametric Bayesian mixture models of Gaussian distributions. that directly estimate the number of cell populations. All these methods, except those of Lin et al. [2013], of Cron et al. [2013] and of Dundar et al. [2014], were evaluated in the FlowCAP-I Challenge whose results are presented in Aghaeepour et al. [2013].

However, there is still room for improvement, especially in the definition of the number of cell population and the identification of rare cell populations. In addition, most of those previous approaches have been proposed for a single sample analysis, except for Cron et al. [2013] who recently proposed to use hierarchical Dirichlet process mixture (DPM) of Gaussian distribution models to analyze multiple samples simultaneously. Yet in the case of repeated measurements of flow cytometry data, it can be useful to perform analysis as the samples are acquired (samples are often collected across several time points in a population of patients, for instance included in a clinical trial). In such a case, one would want to use previously acquired sample as informative prior information in the analysis of a new sample. The approach proposed in this paper includes a strategy of se-

quential approximations of the posterior distribution for multiple data samples, presented in section 4.4.3.

The automated recognition of cell population from flow cytometry data is a difficult task which can be seen as an unsupervised clustering problem. It is characterized by two big challenges. First, the total number of cell populations to identify is unknown. Second, the empirical distributions of the populations are heavily skewed, even when optimal transformation of the data is applied [Pyne et al., 2009; Finak et al., 2010], and the data generally present many outliers. To address all these points together, our approach consider a Bayesian nonparametric model-based approach, where the flow cytometry data are assumed to be drawn from a DPM skew t-distributions. First, this approach enables the number of cell populations to be inferred from the data, and avoids the challenging problem of model selection. Second, it has been demonstrated that the Gaussian assumption for the parametric shape of a cell population fits poorly flow cytometry data [Mosmann et al., 2014]. Indeed, even after state-of-the-art transformation of raw cytometry data, such as the biexponential transformation [Finak et al., 2010], cell population distributions are typically skewed. Pyne et al. [2009] have showed the advantages of the skew t-distribution [Azzalini and Capitanio, 2003] for modeling cell subpopulations in flow cytometry data. The skew t-distribution is a generalization of the skewed normal distribution, with a heavier tale which makes it more robust to outliers. Frühwirth-Schnatter and Pyne [2010] proposed a finite mixture model of skew t-distributions. We extend this model to the infinite mixture case in a Bayesian non parametric framework. Of interest, quantifying the uncertainty around the estimated partition is straightforward in this Bayesian paradigm, from the posterior distribution of the partition. Furthermore, the use of a Bayesian framework allow the use of informative priors. In the case of repeated measurements for instance, we propose to sequentially estimate the posterior partition of flow cytometry using posterior information from time point t as prior information for time point $t + 1$.

4.4.1 A statistical model of Dirichlet process mixture of skew t-distributions

In this section, only one single dataset is considered. The case of the sequential estimation of multiple datasets will be addressed in section 4.4.3. Typically, the data y_c have been transformed from the raw data of measured fluorescence through a biexponential or Box-Cox transformation [Finak et al., 2010].

Multivariate skew t-distribution

First let's consider the choice of the parametric density f_θ which is a skew t-distribution.

Skew normal distribution Frühwirth-Schnatter and Pyne [2010] present a parametrization of the multivariate skew Normal distribution defined by Azzalini and Valle [1996] which leads to the following probability density function:

$$f_{\mathcal{SN}}(y; \xi, \Omega, \eta) = 2\phi(y - \xi; \Omega)\Phi(\eta'\omega^{-1}(y - \xi)) \tag{18}$$

with $\phi(\cdot; \Omega)$ the probability density function of the multivariate Normal distribution with zero mean $\mathcal{N}(0, \Omega)$ and $\Phi(\cdot)$ the cumulative density function of the standard univariate Normal distribution $\mathcal{N}(0, 1)$.

Frühwirth-Schnatter and Pyne [2010] propose a random-effects model representation of such a skew Normal distribution, with truncated normal random effects:

$$Y = \xi + \psi Z + \varepsilon \tag{19}$$

with $Z \sim \mathcal{N}_{[0;+\infty[}(0, 1)$ a truncated univariate standard Normal distribution and $\varepsilon \sim \mathcal{N}(0, \Sigma)$ a multivariate Normal distribution with zero mean. The original parameters can be recovered from:

$$\Omega = \Sigma + \psi\psi', \quad \eta = \frac{1}{\sqrt{1 - \psi'\Omega^{-1}\psi}}\omega\Omega^{-1}\psi \tag{20}$$

The skew t-distribution Let $X \sim \mathcal{SN}(0, \Omega, \eta)$ and $W \sim \text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2})$. If Y has the following stochastic representation:

$$Y = \xi + \frac{1}{\sqrt{W}}X \tag{21}$$

then it follows a multivariate skew t-distribution $Y \sim \mathcal{ST}(\xi, \Omega, \eta, \nu)$ [Azzalini and Capitanio, 2003]. Equation (21) can be expressed as the following random effect model

$$Y = \xi + \psi\frac{Z}{\sqrt{W}} + \frac{\epsilon}{\sqrt{W}} \tag{22}$$

Following the same parametrization as Frühwirth-Schnatter and Pyne [2010], we write the density of a multivariate skew t-distibution as:

$$f_{\mathcal{ST}}(y; \xi, \Omega, \eta, \nu) = 2f_{\mathcal{T}}(y; \xi, \Omega, \nu) \tag{23}$$

$$\times T_{\nu+d}\left(\eta'\omega^{-1}(y - \xi)\sqrt{\frac{\nu + d}{\nu + Q_y}}\right)$$

with $\omega = \sqrt{Diag(\Omega)}$, $Q_y = (y-\xi)'\Omega^{-1}(y-\xi)$, f_T the multivariate Student t-distribution probability density function, and T_ν the cumulative distribution function of the scalar standard Student t-distribution with ν degrees of freedom. Figure 4.4 shows an example of such distributions, highlighting the skewness of both the skew Normal and the skew t and the heavier tail of the skew t distribution.
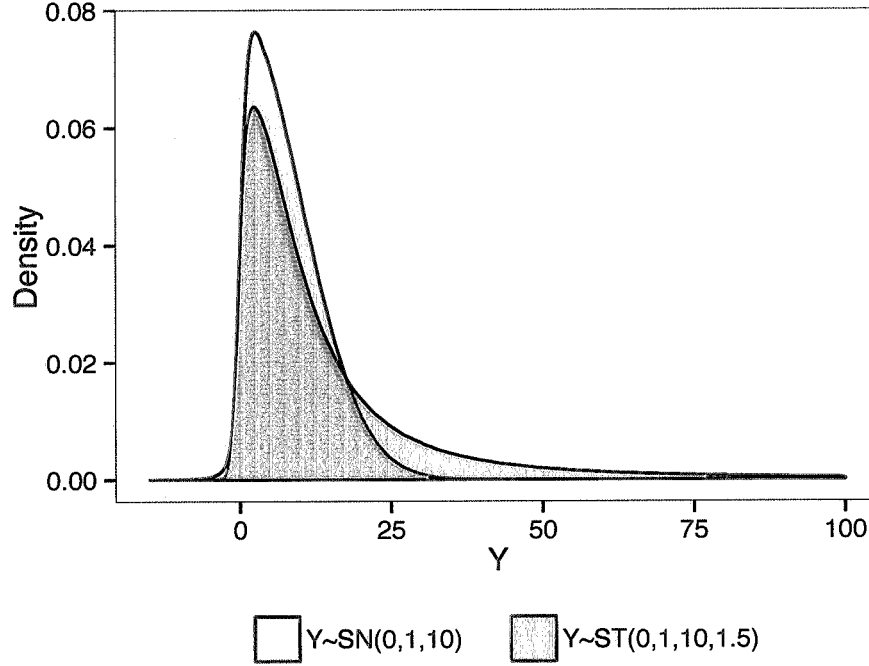
Figure 4.4 – Density probability function of univariate skew Normal $\mathcal{SN}(\xi = 0, \psi = 10, \sigma = 1)$ and skew t $\mathcal{SN}(\xi = 0, \psi = 10, \sigma = 1, \nu = 1.5)$ distributions

4.4.2 Dirichlet process mixture of skew t-distributions

Let G_0 be the base distribution of a Dirichlet process in a DPM combining model (17) with a random-effects model representation (22) of the skew t-distribution. G_0 is the product of a structured inverse Wishart and of a prior on ν, the degree of freedom of the skew t: $G_0 = sNiW(\xi_0, \psi_0, B_0, \Lambda_0, \lambda_0)P_{0,\nu}$. Our proposed model is fully written as follows:

$$\alpha|a,b \sim \text{Gamma}(a,b) \tag{24a}$$

$$\pi \mid \alpha \sim \text{GEM}(\alpha) \tag{24b}$$

for $k = 1, 2, \ldots$

$$\xi_k, \psi_k, \Sigma_k, \nu_k \sim G_0 \tag{24c}$$

for $c = 1, 2, \ldots, C$

$$\ell_c \mid \pi \sim \text{Mult}(\pi) \tag{24d}$$

$$\gamma_c \mid \ell_c, (\nu_k) \sim \text{Gamma}\left(\frac{\nu_{\ell_c}}{2}, \frac{\nu_{\ell_c}}{2}\right) \tag{24e}$$

$$s_c \mid \gamma_c \sim \mathcal{N}_{[0,+\infty[}\left(0, \frac{1}{\gamma_c}\right) \tag{24f}$$

$$y_c \mid \ell_c, \gamma_c, s_c, (\xi_k, \psi_k, \Sigma_k) \sim \mathcal{N}\left(\xi_{\ell_c} + \psi_{\ell_c} s_c, \frac{1}{\gamma_c}\Sigma_{\ell_c}\right) \tag{24g}$$

Discussion on the model assumptions

In model (24), the base distribution parameter G_0 conveys the prior information on the cluster parametric shape. For the parameters ξ_k, ψ_k and Σ_k, we have conditional conjugacy with the random-effects model representation using joint priors taking the form of a structured Normal-inverse-Wishart distribution[Frühwirth-Schnatter and Pyne, 2010]. See Appendix F page 157 for details. Frühwirth-Schnatter and Pyne [2010] pointed out that the prior on Σ_k can have a big impact on the posterior number of clusters. Indeed, setting the scale of the prior on Σ_k too small will give in an inflated number of clusters in the posterior, whereas too large values tend to regroup all the observations together. Adding a Wishart hyperprior on Σ_k, that carries on conjugacy with the inverse Wishart, enables to relax this impact of the prior [Frühwirth-Schnatter and Pyne, 2010; Huang and Wand, 2013]. Assuming prior independence between each ν_k and also from the three aforementioned parameters, we can use any of the three priors proposed in Juárez and Steel [2010] for instance (such as an objective Jeffrey's prior).

4.4.3 Statistical inference for a Dirichlet process mixture of skew t-distributions model

Posterior inference via Gibbs sampling

For making inference on the model (24), MCMC methods allows to sample the partition $\{\ell_{1:C}\}$ as well as the corresponding cluster parameters $\{\theta_k^*\} = \{\{\xi_k^*\}, \{\psi_k^*\}, \{\Sigma_k^*\}, \{\nu_k^*\}\}$ from the marginal posterior distribution. Combining results from Frühwirth-Schnatter and Pyne [2010] and Caron et al. [2014], it is possible to implement an efficient and valid

partially collapsed Gibbs sampler with an Metropolis-Hastings step [van Dyk and Park, 2008; van Dyk and Jiao, 2014]. The use of slice sampling (the idea of sampling from a distribution by uniformly sampling points uner its probability density curve) [Neal, 2003; Walker, 2007; Kalli et al., 2011] enables the straightforward parallelization of the latent allocation sampling (thanks to conditional conjugacy) in such an MCMC algorithm (even in the skew normal and skew t cases). This can lead to substantial computation speed up as the number of observations C (cells) per sample increases. Each iteration of our Gibbs sampler proceeds in the following order (details are provided in supplementary material, see Appendix F page 157):

1. Update the concentration parameter α given the previous partition $\{\ell_{1:C}\}$ using the data augmentation technique from Escobar and West [1995].

2. Update the mixing distribution G given α, $\{\xi_k\}$, $\{\psi_k\}$, $\{\Sigma_k\}$ and the base distribution G_0 via slice sampling.

3. For $c = 1, \ldots, C$ update the individual skew parameter s_c given $\{\xi_k\}$, $\{\psi_k\}$, $\{\Sigma_k\}$ and the new ℓ_c.

4. Update $\{\xi_k\}$, $\{\psi_k\}$, $\{\Sigma_k\}$ given the base distribution G_0, the updated partition $\{\ell_{1:C}\}$ and the updated individual skew parameters $\{s_{1:C}\}$.

5. Finally jointly update the degrees of freedom and the individual scale factors ($\{\nu_k\}$, $\{\gamma_{1:C}\}$) in an Metropolis-Hastings (M-H) within Gibbs step. First an M-H step is performed to update the $\{\nu_k\}$ where the $\{\gamma_{1:C}\}$ are integrated out, immediately followed by a Gibbs step to sample the $\{\gamma_{1:C}\}$ from their full conditional distribution. This ensures that the reduced conditioning performed in the M-H step does not change the stationary distribution of the Markov chain [van Dyk and Jiao, 2014] – see Appendix F page 157.

Sequential Posterior approximation

In flow cytometry experiments, it is common to actually have multiple datasets $y^{(i)}$ (with $i = 1, \ldots, I$) corresponding to multiple individuals, or repeated measurements of the same individual. For instance in the case of clinical trials, longitudinal measurements of flow cytometry data are often performed for the same patients. In such cases, it is of interest to use previous time points or previous samples results as prior information. Specifying prior information to Dirichlet process mixture models not straightforward [Kessler et al., 2015]. We propose to use the posterior MCMC draws obtained from previous

dataset $y^{(i)}$ as prior information to analyze the next dataset $y^{(i+1)}$:

$$\alpha \sim Gamma(a, b) \tag{25a}$$

$$G|\alpha \sim DP(\alpha, G_0) \tag{25b}$$

$$y^{(i)}, y^{(i+1)}|G \sim^{iid} \int_{\Theta} f_{\theta}(\cdot)dG(\theta) \tag{25c}$$

We are interested in estimating $p(G|y^{(i)}, y^{(i+1)}) \propto p(G|y^{(i)})p(y^{(i+1)}|G)$. The idea is to first approximate $p(G|y^{(i)})$ by a Dirichlet process through MCMC draws from the model described in 4.4.1:

$$p(G|y^{(i)}) \simeq \int DP(G; \alpha, G_1)Gamma(\alpha; a_1, b_1)d\alpha \tag{26}$$

where G_1, a_1, b_1 are parameters to be estimated from the MCMC approximation of the true posterior: i) $\widehat{a_1}$ and $\widehat{b_1}$ can be taken as MLE estimates from the MCMC samples $\alpha^{(j)}$; ii) $\widehat{G_1}$ is a parametric approximation of the posterior. Indeed, the posterior of G_1 is not suitable for being directly plugged in as a base distribution parameter of another DP. In the case of a skew t-distributions mixture, G_1 is a joint distribution: $G_1 = (sNiW, P_{0,\nu})$ where $P_{0,\nu}$ is the chosen prior for the skew t-distribution degrees of freedom. To estimate G_1, we estimate the Maximum *a posteriori* (MAP) from the posterior MCMC samples (Appendix G page 163).

Using this posterior parametric approximation, we have the same hierarchical model, conditional on $y^{(i)}$:

$$\alpha|y^{(i)} \sim Gamma(\widehat{a_1}, \widehat{b_1}) \tag{27a}$$

$$G|\alpha, y^{(i)} \sim DP(\alpha, \widehat{G_1}) \tag{27b}$$

$$y^{(i+1)}|G, y^{(i)} \sim^{iid} \int_{\Theta} f_{\theta}(\cdot)dG(\theta) \tag{27c}$$

Note that under this approximate posterior model, the cluster parameters $\theta_j^{(i)}$ are *iid* from G_1. Such an approach can be iterated numerous time if for instance several time points are observed, approximating the successive posteriors.

Point estimate of the cell populations

Getting a representation of the partition posterior distribution is difficult. One can use the maximum a posteriori, i.e. using the point estimation form the MCMC sample that

maximize the posterior density. This approach loses any sense of the uncertainty conveyed by the Bayesian approach. We rather consider a co-clustering probability matrix ζ on each pair (k, l) of observations. Such a matrix can be estimated by averaging the co-clustering matrices from all the explored partitions in the posterior MCMC draws:

$$\widehat{\zeta}_{cd} = \frac{1}{N} \sum_{i=1}^{N} \delta_{\ell_c^{(i)} \ell_d^{(i)}} \tag{28}$$

where N is the number of MCMC draws from the posterior and $\delta_{kl} = 1$ if $k = l$, 0 otherwise. The computational cost of this approach, though, is of the order $\mathcal{O}(Nn^2)$.

An optimal partition point estimate $\{\widehat{\ell}_{1:C}\}$ can then be derived in regard of this similarity matrix by using a pairwise coincidence loss function [Lau and Green, 2007], such as the one proposed by Binder [1978, 1981]:

$$\{\widehat{\ell}_{1:C}\} = \underset{\{\ell_{1:C}^{(i)}\} \in \left\{\{\ell_{1:C}^{(1)}\}, \dots, \{\ell_{1:C}^{(N)}\}\right\}}{\arg\min} \sum_{c=1}^{C-1} \sum_{d=c+1}^{C} 2 \left(\delta_{\ell_c^{(i)} \ell_d^{(i)}} - \widehat{\zeta}_{cd}\right)^2 \tag{29}$$

An optimal partition point estimate $\{\widetilde{\ell}_{1:C}\}$ can also be derived in regard of the F-measure. The F-measure is widely used as a way to summarize the accordance between 2 methods, one being considered as a reference gold-standard. It is the harmonic mean of the precision and recall:

$$F = \frac{2PrRe}{Pr + Re} \tag{30}$$

In order to use the F-measure to evaluate our clustering method, we rely on the definition proposed in the online methods from Aghaeepour et al. [2013]. In this setting of unsupervised clustering, the precision Pr is the number of cells correctly assigned to a given cluster divided by the total number of cells assigned to this cluster. It can also be called Positive Predictive Value. The recall Re is the number of cells correctly assigned to a given cluster divided by the number of cells that should be assigned to this cluster according to the gold-standard. Since in our problem the labels of the different clusters are exchangeable, the F-measure is computed for each combination of the reference clusters and the predicted clusters. Let $G = \{g_1, \dots, g_m\}$ be a set of m reference clusters and $H = \{h_1, \dots, h_n\}$ be set of n predicted clusters. For each combination pair (q, r) of a reference cluster g_q and a predicted cluster h_r, the F-measure is computed as follows:

$$Pr(h_r, g_q) = \frac{|g_q \cap h_r|}{|h_r|} \quad \text{and} \quad Pr(h_r, g_q) = \frac{|g_q \cap h_r|}{|g_q|} \tag{31}$$

$$F(h_r, g_q) = \frac{2Pr(g_q, h_r)Re(g_q, h_r)}{Pr(g_q, h_r) + Re(g_q, h_r)} \tag{32}$$

This F-measure is comprised in $[0, 1]$, the closer it is to 1 the better agreement between the predicted cluster and the reference cluster. The total F-measure for a predicted partition

H given a gold-standard G is then define as the weighted sum of the best matched F-measure:

$$F_{tot}(H, G) = \frac{1}{\sum_{q=1}^{m} |g_q|} \sum_{q=1}^{m} |g_q| \max_{r \in \{1,\ldots,n\}} F(h_r, g_q) \tag{33}$$

This total F-measure is comprised in $[0, 1]$, and the closer to 1 it gets, the closer the predicted partition is from the gold-standard partition. The optimal partition point estimate in regard of this thus defined F-measure can be derived by maximizing the average F-measure over all the explored partitions in the posterior MCMC draws:

$$\{\tilde{\ell}_{1:C}\} = \underset{\{\ell_{1:C}^{(i)}\} \in \{\{\ell_{1:C}^{(1)}\},\ldots,\{\ell_{1:C}^{(N)}\}\}}{\arg\max} \frac{1}{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} F_{tot}\left(\{\ell_{1:C}^{(i)}\}, \{\ell_{1:C}^{(j)}\}\right) \tag{34}$$

4.4.4 Applications

Simulations

Two different simulations studies were conducted: first to evaluate the performance of the Dirichlet process mixture of skew t distributions model in a simple clustering case, and second to quantify the improvement of the sequential posterior approximation model over the broad prior strategy in a more realistic scenario.

To assess the proposed model, 100 simulations in 2-dimensions were performed. 2000 observations were simulated. Three scenarios were used, in which observations were simulated from: i) the model thanks to the Chinese Restaurant process; ii) 4 clusters easily separated clusters; iii) 4 overlapping clusters. inn the last two scenarios, the four cluster represented respectively 50%, 30%, 15% and 5% of the data. After 10,000 MCMC iterations (8,000 iterations burnt and a thining of 10 gave 200 partitions sampled from the posterior; the chain was initialized with 30 clusters), the resulting mean F-measure when comparing the point estimate obtained from our approach with the true original clustering of the simulated data were, in the three scenarios respectively, 0.943, 0.999, and 0.870. Figure 4.5 shows an example of the partition point estimate obtained for one of those 100 simulations in both the "easy" and the "overlapping" scenarios.

To evaluate the sequential posterior approximation model, 100 simulations in 2-dimensions were performed. For each simulation, two datasets were simulated, each of 2000 observations. First a learning simulated sample with 4 distinct clusters following skew t-distributions, representing respectively 70%, 10%, 10% and 10% of the data. Second a test simulated sample with 3 distinct clusters representing respectively 80%, 10% and 10% of the data (Figures 4.6A. and 4.6B. for an example). The large cluster in the second sample was actually the superposition of 3 different skew t-distributions at the same locations

A: "easy" scenario
B: "overlapping" scenario
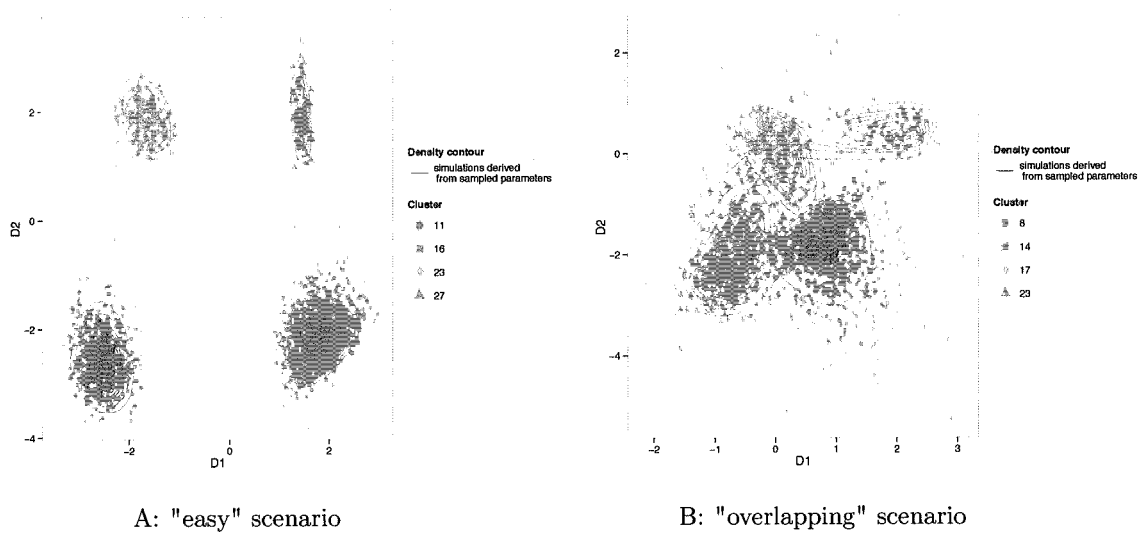
Figure 4.5 – Partition point estimate from one of the 100 2-dimensional simulations

as the lower left cluster of the learning sample. After 10,000 MCMC iterations (8,000 iterations burnt and a thining of 10 gave 200 partitions sampled from the posterior; the chain was initialized with 30 clusters) the resulting mean F-measure was of 0.904 with a broad prior but was improved to 0.983 with the sequential posterior approximation when comparing the point estimate obtained from our approach with the true original clustering of the simulated data. Figures 4.6C. and 4.6D. show an example of the partition point estimate obtained for one of the 100 simulations. The informative prior lead to more robust inference in this case: adding prior knowledge from the learning sample that there is one large lower-left cluster enabled the model to fit only one cluster there (Figures 4.6C., 4.6D.). The posterior partition of the data shows much less variability when an informative prior is used (Figures 4.6E. and 4.6F). However, the prior knowledge that there was one cluster in the lower right from the learning sample was not backed by the data from the test sample, and the model correctly estimated no cluster located there (in spite of this prior information). So the informative prior strategy seems to give more robust estimations, for instance when the clusters are not stricly skew t-distributed.

Benchmark real experimental datasets analyses

Two real experimental datasets are analyzed with the proposed approach. Both were used as benchmark data in Aghaeepour et al. [2013]. First the Graft versus Host Disease (GvHD) dataset, a public dataset was first analysed (manually gated) in Brinkman et al. [2007], with the objective of identifying cellular signature that correlates or predict Graft versus Host disease. Flow-cytometry data was collected for 12 samples. Second the Hematopoietic Stem Cell Transplant (HSCT) dataset, which consists of 30 samples of

A. Simulated learning sample data

B. Simulated test sample data

C. Partition point estimate of the test sample data
with a broad prior

D. Partition point estimate with an informative prior

E. Similarity matrix of the test sample data
clustering with a broad prior

F. Similarity matrix of the test sample data
clustering with an informative prior
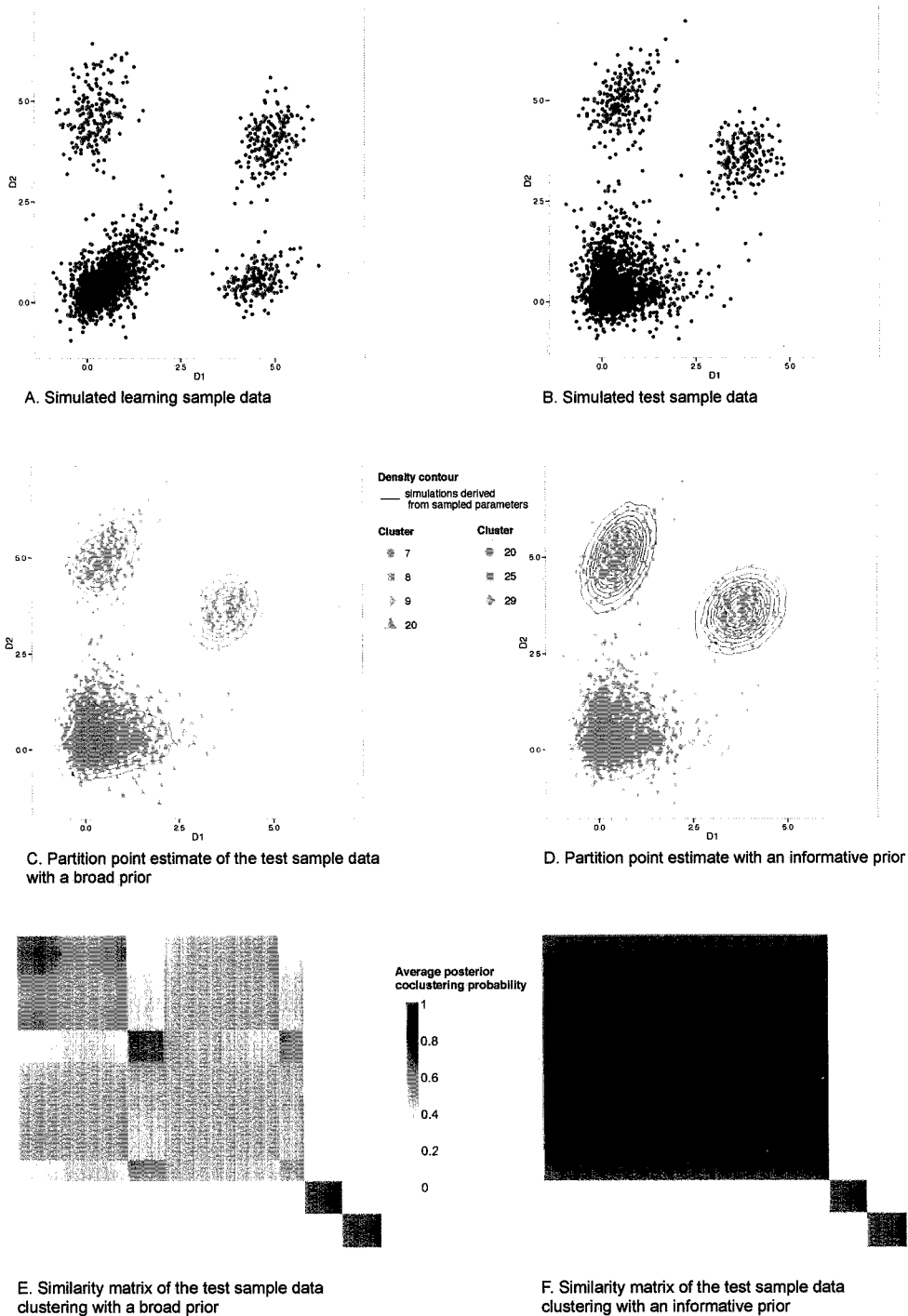
Figure 4.6 – Example from one of the second 100 2-dimensional simulations

bone marrow cells from the Terry Fox laboratory. For both datasets, their original manual gates are being regarded as the true cell clustering, despite the reproducibility issues with manual gating [Ge and Sealfon, 2012; Aghaeepour et al., 2013]. In order to compare our method, we used the exact same data as in Aghaeepour et al. [2013], provided by the FlowCAP project [http://flowcap.flowsite.org/] as part of the FLowCAP-I challenge [http://ficcs.org/dataFiles/FlowCAP-I.zip]. Table 4.1 shows the performance of our proposed approach NPflow on these two datasets, in the context of the other approaches reviewed by Aghaeepour et al. [2013]. The F-measure is computed for all samples available for a given dataset and the mean over all samples is reported. The only algorithms performing significantly better than NPflow in both benchmark datasets are flowMeans [Aghaeepour et al., 2011] and FLAME [Pyne et al., 2009].

It is difficult to interpret the result of the sequential posterior approximation model when applied to any of those 2 experimental datasets because to the lack of clinical information. Indeed, in both case samples preceding the transplantation as well as samples after the transplantation are considered. Samples before and after transplantation are supposed to be very different, and thus question the hypothesis behind the sequential posterior model.

4.4.5 Conclusion

We extend the classical Dirichlet process Gaussian mixture model to skew t-distribution mixtures, based on Frühwirth-Schnatter and Pyne [2010] parametrization of such distributions. Such a model is especially suited for model based unsupervised classification of flow cytometry data. Automated gating of cell populations is an open research problem and the proposed approach features three important characteristics for this task: i) it avoids the difficult issue of model selection by estimating directly the number of component in the mixture ; ii) it uses skew and heavy tailed distributions in the form of skew t-distributions, of which the gaussian is a particular case, iii) it provides estimation of the posterior co-clustering probabilities for each data pair that allows to quantify the uncertainty about the partition. An efficient collapsed Metropolis within Gibbs sampler has been developed for estimating such models. It has been implemented in ℝ as a package NPflow, which is still under development but will soon be made available to the community.

Some gain in computation time can be obtained through CPU parallelization and C++ implementation of key bottleneck within the sampler. Yet, computation times for such models remains important. Depending on the number of data and the number of clusters it ranges from a few minutes up to a couple hours (for around 100,000 observations). Since

Method	GvHD (n=12)	HSCT (n=30)
NPflow	0.75 (0.70, 0.81)	0.82 (0.79, 0.85)
ADICyt	0.81 (0.72, 0.88)	0.93 (0.90, 0.96) [+]
CDP	0.52 (0.46, 0.58) [−]	0.50 (0.48, 0.52) [−]
FLAME	0.85 (0.77, 0.91) [+]	0.94 (0.92, 0.95) [+]
FLOCK	0.84 (0.76, 0.90)	0.86 (0.83, 0.89)
flowClust/Merge	0.69 (0.55, 0.79)	0.81 (0.77, 0.85)
flowMeans	0.88 (0.82, 0.93) [+]	0.92 (0.90, 0.94) [+]
FlowVB	0.85 (0.79, 0.91) [+]	0.75 (0.70, 0.79) [−]
L2kmeans	0.64 (0.57, 0.72) [−]	0.70 (0.65, 0.75) [−]
MM	0.83 (0.74, 0.91) [+]	0.73 (0.66, 0.80)
MMPCA	0.84 (0.74, 0.93)	0.91 (0.88, 0.94) [+]
SamSPECTRAL	0.87 (0.81, 0.93) [+]	0.85 (0.82, 0.88)
SWIFT	0.63 (0.56, 0.70) [−]	0.59 (0.55, 0.62) [−]

All estimates except for our proposed NPflow approach are from Aghaeepour et al. [2013]. 95% Confidence Intervals are calculated on 1,000 bootstrap samples of the F-measures. A [−] denotes the methods significantly worse than the proposed NPflow model (according to a paired signed rank test at a 0.05 level), and a [+] denotes the methods with significantly better, in regards of the manual gating reference.

Table 4.1 – Mean F-measures across all samples on two benchmark experimental datasets

parallelized operation are very short, the gain in computation time is quickly overtaken by time lost in communication between CPUs as their number increase.

In case of repeated measurement of flow cytometry data, such as in a clinical trial, the proposed sequential analysis strategy enables to analyze each sample sequentially, as the data are acquired. It requires neither to wait for the last sample to perform the automated gating nor to analyze all data at once, but it still uses available prior knowledge. This contrasts with hierarchical extensions of the Dirichlet Process Mixture Model such as those proposed by Cron et al. [2013] or Dundar et al. [2014], where the complete dataset must be analyzed at once. In our simulation study this sequential prior strategy improves the fit of the model.

Manual gating is considered as the gold-standard when evaluating an automated gating strategy on real flow cytometry data. Yet one should keep in mind that manual gating has reproducibility issues, often resulting in a partial and subjective clustering [Ge and Sealfon, 2012; Aghaeepour et al., 2013]. Therefore manual gating might not be actually

the ideal way to assess the performance of automated gating algorithms, and consensus clustering of manual or automated operators could be used instead.

5 General discussion

This thesis covers various aspects of real world data encountered today in vaccine trials. Thank to the technological improvements, the available data are now high-dimensional and, above all, complex. In this work, we tried to present new tools to analyze such data, and described a few ways to integrate different biological levels in a single analysis. The key idea is to use the maximum information available. This is attempted in two ways: i) the use of prior knowledge, that requires to transform expertise into mathematical properties; ii) the use of all data at hand, hopefully in an integrated manner. We believe that this addition of meaningful information will ease the separation of relevant signal from noise, and will make the inference of interpretable biological results from high-dimensional data easier and more consistent.

Longitudinal gene expression measurements are becoming more and more common in clinical trials thanks to the decreasing cost of microarray technology. Such omics data are prone to technical variability, and must analyzed carefully in order to differentiate signal from noise. However, a lot of biological knowledge on gene pathways already exists. It can be leveraged in order to enhance the sensitivity and the interpretability of such an analysis. Taking into account this prior knowledge, we proposed to harness the power linear mixed modeling coupled with a flexible control of the false discovery rate to model gene expression over time. This modeling allowed us to unravel the impact of a therapeutic vaccine against HIV on the state of gene expression in the blood of the vaccinated patients. The signature thus highlighted had some overlap with signature derived in other vaccines (namely the pneumococcal and flu vaccines). Compared to other methods available in the literature [Subramanian et al., 2005; Efron and Tibshirani, 2007; Hummel et al., 2008; Shahbaba et al., 2011; Wu and Smyth, 2012], we have underlined the usefulness of our approach to answer the key questions: "Which gene sets have expression dynamics moving significantly over time ?".

In gene set analyses, the gene sets' definition is of utmost importance. Indeed, regardless of the method (self-contained or competitive), a gene set analysis can be viewed as hypothesis driven. The definition of the gene sets tested is therefore constitutive of the hypotheses tested. In this thesis, two different strategies for gene sets definition are considered: i) *a priori* defined gene sets (the Modules from Chaussabel et al. [2008]) in the DALIA-1 trial; ii) gene sets directly inferred from the data itself [Furman et al., 2013]

in the flu vaccine systems analysis. However, we did not tackle the issue of gene network inference [Marbach et al., 2012; de la Fuente, 2013]. Gene network inference aims at increasing biological knowledge on gene pathways from genomic data. It is usually performed on time-course gene expression data. Recently, Wu et al. [2014] proposed sparse ordinary differential equation modeling for gene network reconstruction from time-course expression data in a very promising first attempt to adapt differential equation modeling to high-dimensional data. Further developments are needed here to improve the identifiability of parameters in such complex modeling approach. Other approaches such as the Approximate Bayes Computation could be a good alternative [Ratmann et al., 2009].

RNA-seq is another technology to measure gene expression. It gives a more precise measurement of the expression compared to microarrays, and it is increasingly used in biological and medical studies. However, as a result of the increased precision, RNA-seq data are count data. This makes modeling RNA-seq data more difficult, in particular, it is not clear which statistical distribution is best suited for modeling such data (over-dispersed Poisson distribution as well as the negative binomial distribution have been proposed for instance [Auer and Doerge, 2011; Srivastava and Chen, 2010]). Instead, Law et al. [2014] proposed to model logarithm transformed counts of RNA-seq reads with Normal distribution while taking into accounts heteroskedasticity with in precision weights. They argue that correct modeling of variance is the key to powerful statistical tools of analysis. Such methodology would be directly applicable in TcGSA by including precision weights in the mixed models. Consequently, our approach is easily extendable to RNA-seq data.

But the longitudinal gene expression was not the only measurement available in this particular therapeutic HIV vaccine trial. Along with it several other phenotypic variables were measured including cytokine production, cell functionality, etc. Starting from the signature identified with our proposed approach for analyzing time-course gene expression, we related gene expression to those phenotypic data thanks to sparse partial least squares. It highlighted an association of low inflammatory pathway expression with lower viral rebound. This association was consistent through a sensitivity analysis. In a different vaccine trial (against influenza), more traditional analyses revealed an association between a testosterone modulated group of gene and the higher immune response to the vaccine in females compared to males. In this study, high-dimensionality of the data was dealt with by first aggregating gene expressions into gene modules, and then by correcting for multiple testing. In both cases, integration of various data types cast a light on potential underlying biological explanations of the heterogeneity among the patients' vaccine response.

Finally, we developed a state-of-the-art nonparametric bayesian clustering was with

a specific focus on automated processing of flow cytometry data. We proposed to extend the classical Gaussian Dirichlet process mixture model to the more general skew t-distributions. Skew t-distribution are a class of probability distributions that are potentially asymmetric and heavy tailed, of which the normal distribution is a particular case. Given the shape of raw flow-cytometry fluorescence data, such properties are attractive. We also proposed an efficient partially collapsed Metropolis within Gibbs sampler for estimating such models. Such a model performs similarly as other approach developed for automated clustering of flow cytometry data. It has the advantage allowing the use of informative prior specifications, in particular on the cluster locations. Such informative priors can be used for instance in sequential posterior approximation for instance to take advantage of repeated measurements. We expect in the near future that this type of approach could be broadly used by immunologists who really need it to face to increasing dimension of flow cytometry data.

Cron et al. [2013] proposed a hierarchical Gaussian Dirichlet process mixture model closely related to our approach. They model between-sample heterogeneity by introducing a hierarchical layer in their model. Dundar et al. [2014] proposed a similar model to identify outlier samples. Both models analyze all data simultaneously. Cron et al. [2013] uses GPU parallelization in order to rapidly compute MCMC estimations of their hierarchical Gaussian Dirichlet process mixture model on such a big amount of data. Our approach takes a different approach by dealing with several samples sequentially. It is currently implemented through CPU parallelization would likely benefit from switching to GPU parallelization, which is more efficient when rapid computations have to be repeated a large number of times (such as in the cluster allocation step in our gibbs sampler, see Appendix F page 157). However, as all dimensions (number of samples, number of markers, number of cells) of flow-cytometry data increase, online data modeling (where new observations become sequentially available) is appealing, especially in longitudinal studies. In this context, other methods for estimating Dirichlet process mixtures models such as Sequential Monte Carlo (SMC) samplers (or particle filters) [Doucet et al., 2001; Del Moral et al., 2006] offer promising properties for dynamic modeling. Indeed, such algorithms are naturally online and easily parallelized.

In the context of automated gating of flow cytometry data, manual gating is considered as the gold-standard. Of course, it is important that automated approaches exhibit good performances compared to manual gating in order to convince immunologists that they are good alternatives. However it has been shown that manual gating, among other issues, is highly variable from one operator to another [Ge and Sealfon, 2012; Aghaeepour et al., 2013]. This makes it difficult to assess automated gating algorithms because the truth is unknown. In addition, due to the increase in the number of markers measured

simultaneously, it now becomes impossible to manually look at all possible marker combinations in a manual gating strategy, whereas automated algorithms usually take into account all dimensions. Simulating realistic flow-cytometry data is difficult, as those are highly structured, sparse, high-dimensional data [Finak et al., 2009; Pyne et al., 2009]. To overcome this obstacle, one could use consensus clustering, either among automated gating algorithms, like Aghaeepour et al. [2013], or among technical operator performing manual gating (if several are available).

One of the perspective of automated gating of flow cytometry data is to improve the reproducibility of the final subpopulation cell counts. This goal requires an extra step after the gating of the cell population (regardless of wether a manual or automated method was used): the annotation of the cell clusters. This task shares some similarities with gene sets analyses, although the two are quite different. Very recently, Courtot et al. [2014] proposed an algorithm using cell ontology [Diehl et al., 2011] to automatically annotate cell populations based on a dichotomous split of each cellular marker. This next step of the automated processing of flow cytometry data seems to be a new research area, whose importance will likely increase as automated gating methods are becoming increasingly performant.

Both the approach for the gene set analysis of longitudinal gene expression (TcGSA) and the Dirichlet process mixture of skew t-distributions model have been implemented as ⓡ packages: TcGSA v0.9.8 is available on CRAN, whereas NPflow is still under development (but will be made available to the community). Today, dissemination of statistical developments requires their implementation in widely used softwares such as ⓡ.

A next step for data integration would be to use the deconvolution in the spirit of Shen-Orr et al. [2010] but for repeated measurements of gene expression (gene abundance) and cell populations. It would be then of interest to disentangle the reason of changes in gene abundance that can be due either to a real modification of the gene expression at the cell level or to the circulation of some some specific cell populations.

The analysis of all of these data requires new methods that are still under development and that need to be adapted according to the question asked, and consequently to the data available. These complex data constitute the observations of a complex system, that require sophisticated modeling, and preclude the use of intuitive simplistic statistics. This opens a new area where biostatisticians has never been as much demanded, but where collaborations have never been as much needed.

Bibliography

Abdool Karim Q., Abdool Karim S. S., Frohlich J. A., Grobler A. C., Baxter C., Mansoor L. E., Kharsany A. B. M., Sibeko S., Mlisana K. P., Omar Z., Gengiah T. N., Maarschalk S., Arulappan N., Mlotshwa M., Morris L., and Taylor D. Effectiveness and safety of tenofovir gel, an antiretroviral microbicide, for the prevention of HIV infection in women. *Science*, 329(5996): 1168–74, 2010. 32

Ackermann M. and Strimmer K. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10: 47, 2009. 18, 44, 46

Aghaeepour N., Nikolic R., Hoos H. H., and Brinkman R. R. Rapid cell population identification in flow cytometry data. *Cytometry. Part A : the journal of the International Society for Analytical Cytology*, 79(1): 6–13, 2011. 90, 101

Aghaeepour N., Finak G., Hoos H., Mosmann T. R., Brinkman R. R., Gottardo R., and Scheuermann R. H. Critical assessment of automated flow cytometry data analysis techniques. *Nature methods*, 10(3): 228–38, 2013. 25, 26, 89, 90, 97, 99, 101, 102, 107, 108

Airola A., Pahikkala T., Waegeman W., De Baets B., and Salakoski T. An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Computational Statistics & Data Analysis*, 55(4): 1828–1844, 2011. 78

Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 1974. 141

Antoniak C. E. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6): 1152–1174, 1974. 25, 89

Ashburner M., Ball C. A., Blake J. A., Botstein D., Butler H., Cherry J. M., Davis A. P., Dolinski K., Dwight S. S., Eppig J. T., Harris M. A., Hill D. P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J. C., Richardson J. E., Ringwald M., Rubin G. M., and Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1): 25–9, 2000. 16, 33, 42, 43, 61, 125

Auer P. L. and Doerge R. W. A Two-Stage Poisson Model for Testing RNA-Seq Data. *Statistical Applications in Genetics and Molecular Biology*, 10(1): 1–26, 2011. 106

Auvert B., Taljaard D., Lagarde E., Sobngwi-Tambekou J., Sitta R., and Puren A. Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: the ANRS 1265 Trial. *PLoS Medicine*, 2(11): e298, 2005. 32

Azzalini A. and Valle A. D. The multivariate skew-normal distribution. *Biometrika*, 83(4): 715–726, 1996. 92

Azzalini A. and Capitanio A. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2): 367–389, 2003. 25, 91, 92

Barry W. T., Nobel A. B., and Wright F. a. Significance analysis of functional categories in gene expression studies: A structured permutation approach. *Bioinformatics*, 21(9): 1943–1949, 2005. 41

Bécavin C., Tchitchek N., Mintsa-Eya C., Lesne A., and Benecke A. Improving the efficiency of multidimensional scaling in the analysis of high-dimensional data using singular value decomposition. *Bioinformatics*, 27(10): 1413–21, 2011. 43, 53, 54

Berk M., Hemingway C., Levin M., and Montana G. Longitudinal Analysis of Gene Expression Profiles Using Functional Mixed-EffectsModels. In Di Ciaccio A., Coli M., and Angulo Ibanez J. M., editors, *Advanced Statistical Methods for the Analysis of Large Data-Sets*, pages 57–67. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. 43

Berry M. P. R., Graham C. M., McNab F. W., Xu Z., Bloch S. a. a., Oni T., Wilkinson K. a., Banchereau R., Skinner J., Wilkinson R. J., Quinn C., Blankenship D., Dhawan R., Cush J. J., Mejias A., Ramilo O., Kon O. M., Pascual V., Banchereau J., Chaussabel D., and O'Garra A. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature*, 466(7309): 973–7, 2010. 47

Biernacki C., Celeux G., and Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7): 719–725, 2000. 87

Binder D. A. Bayesian Cluster Analysis. *Biometrika*, 65(1): 31, 1978. 97

Binder D. A. Approximations to Bayesian Clustering Rules. *Biometrika*, 68(1): 275, 1981. 97

Boedigheimer M. J., Wolfinger R. D., Bass M. B., Bushel P. R., Chou J. W., Cooper M., Corton J. C., Fostel J., Hester S., Lee J. S., Liu F., Liu J., Qian H.-R., Quackenbush J., Pettit S., and Thompson K. L. Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories. *BMC genomics*, 9: 285, 2008. 40

Bolstad B., Irizarry R. A., Astrand M., and Speed T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2): 185–93, 2003. 39

Bosinger S. E., Jacquelin B., Benecke A., Silvestri G., and Müller-Trutwin M. Systems biology of natural simian immunodeficiency virus infections. *Current opinion in HIV and AIDS*, 7(1): 71–8, 2012. 43, 53, 54

Boulesteix A.-L. and Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in bioinformatics*, 8(1): 32–44, 2007. 131

Brinkman R. R., Gasparetto M., Lee S.-J. J., Ribickas A. J., Perkins J., Janssen W., Smiley R., and Smith C. High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. *Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation*, 13(6): 691–700, 2007. 99

Bůžková P., Lumley T., and Rice K. Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions. *Annals of Human Genetics*, 75: 36–45, 2011. 76

Caron F., Teh Y. W., and Murphy T. B. Bayesian nonparametric Plackett–Luce models for the analysis of preferences for college degree programmes. *The Annals of Applied Statistics*, 8(2): 1145–1181, 2014. 94

Casella G. An introduction to empirical Bayes data analysis. *The American Statistician*, 39(2): 83–87, 1985. 41

Chan C., Feng F., Ottinger J., Foster D., West M., and Kepler T. B. Statistical mixture modeling for cell subtype identification in flow cytometry. *Cytometry. Part A : the journal of the International Society for Analytical Cytology*, 73(8): 693–701, 2008. 90

Chaussabel D., Quinn C., Shen J., Patel P., Glaser C., Baldwin N., Stichweh D., Blankenship D., Li L., Munagala I., Bennett L., Allantaz F., Mejias A., Ardura M., Kaizer E., Monnet L., Allman W., Randall H., Johnson D., Lanier A., Punaro M., Wittkowski K. M., White P., Fay J., Klintmalm G., Ramilo O., Palucka a. K., Banchereau J., and Pascual V. A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity*, 29(1): 150–164, 2008. 18, 21, 23, 42, 43, 47, 53, 54, 57, 61, 105, 125

Chen C., Grennan K., Badner J., Zhang D., Gershon E., Jin L., and Liu C. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PloS one*, 6(2): e17238, 2011. 40

110

Cliff J. M., Lee J.-S., Constantinou N., Cho J.-E., Clark T. G., Ronacher K., King E. C., Lukey P. T., Duncan K., Van Helden P. D., Walzl G., and Dockrell H. M. Distinct phases of blood gene expression pattern through tuberculosis treatment reflect modulation of the humoral immune response. *The Journal of infectious diseases*, 207(1): 18–29, 2013. 47

Cobb A., Roberts L. K., Palucka A. K., Mead H., Montes M., Ranganathan R., Burkeholder S., Finholt J. P., Blankenship D., King B., Sloan L., Harrod A. C., Lévy Y., and Banchereau J. Development of a HIV-1 lipopeptide antigen pulsed therapeutic dendritic cell vaccine. *Journal of Immunological Methods*, 365(1-2): 27–37, 2011. 32

Conesa A., Nueda M. J., Ferrer A., and Talón M. maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*, 22(9): 1096–102, 2006. 44

Cook I. F. Sexual dimorphism of humoral immunity with human vaccines. *Vaccine*, 26(29-30): 3551–3555, 2008. 74

Courtot M., Meskas J., Diehl A. D., Droumeva R., Gottardo R., Jalali A., Renani J. T., Maecker H. T., McCoy J. P., Ruttenberg A., Scheuermann R. H., and Brinkman R. R. flowCL: ontology-based cell population labelling in flow cytometry. *Bioinformatics*, (December 2014): 1–3, 2014. 108

Croft D., Mundo A. F., Haw R., Milacic M., Weiser J., Wu G., Caudy M., Garapati P., Gillespie M., Kamdar M. R., Jassal B., Jupe S., Matthews L., May B., Palatnik S., Rothfels K., Shamovsky V., Song H., Williams M., Birney E., Hermjakob H., Stein L., and D'Eustachio P. The Reactome pathway knowledgebase. *Nucleic Acids Research*, 42(November 2013): 472–477, 2014. 42

Cron A., Gouttefangeas C., Frelinger J., Lin L., Singh S. K., Britten C. M., Welters M. J. P., van der Burg S. H., West M., and Chan C. Hierarchical modeling for rare event detection and cell subset alignment across flow cytometry samples. *PLoS computational biology*, 9(7): e1003130, 2013. 26, 27, 90, 102, 107

Davis M. M. A prescription for human immunology. *Immunity*, 29(6): 835–338, 2008. 73

de la Fuente A., editor. *Gene Network Inference*. Springer, 2013. 106

Del Moral P., Doucet A., and Jasra A. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3): 411–436, 2006. 107

Dempster A., Laird N., and Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–38, 1977. 167

Diehl A. D., Augustine A. D., Blake J. a., Cowell L. G., Gold E. S., Gondré-Lewis T. a., Masci A. M., Meehan T. F., Morel P. a., Nijnik A., Peters B., Pulendran B., Scheuermann R. H., Yao Q. A., Zand M. S., and Mungall C. J. Hematopoietic cell types: prototype for a revised cell ontology. *Journal of biomedical informatics*, 44(1): 75–9, 2011. 108

Diggle P., Heagerty P., Liang K. Y., and Zeger S. *Analysis of longitudinal data*. Oxford University Press, USA, 2002. 46, 61

Doering T. A., Crawford A., Angelosanto J. M., Paley M. A., Ziegler C. G., and Wherry E. J. Network analysis reveals centrally connected genes and pathways involved in CD8+ T cell exhaustion versus memory. *Immunity*, 37(6): 1130–44, 2012. 47

Doucet A., de Freitas N., and Gordon N., editors. *Sequential Monte Carlo Methods in Practice*. Springer, 2001. 107

Dudoit S. and Van der Laan M. J. *Multiple Testing Procedures with Applications to Genomics*. Springer Series in Statistics. Springer (New York), 2008. 49, 122

Dudoit S., Yang Y. H., Callow M. J., and Speed T. P. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica sinica*, 12: 111–139, 2002. 37

Dundar M., Akova F., Yerebakan H. Z., and Rajwa B. A non-parametric Bayesian model for joint cell clustering and cluster matching: identification of anomalous sample phenotypes with random effects. *BMC Bioinformatics*, 15: 314, 2014. 26, 27, 90, 102, 107

Efron B. Two Modeling Strategies for Empirical Bayes Estimation. *Statistical Science*, 29(2): 285–301, 2014. 41

Efron B. and Tibshirani R. On testing the significance of sets of genes. *Annals of Applied Statistics*, 1(1): 107–129, 2007. 18, 27, 41, 43, 44, 105

Escobar M. D. and West M. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430): 577–588, 1995. 25, 89, 95

Fages A., Ferrari P., Monni S., Dossus L., Floegel A., Mode N., Johansson M., Travis R. C., Bamia C., Sánchez-Pérez M.-J., Chiodini P., Boshuizen H. C., Chadeau-Hyam M., Riboli E., Jenab M., and Elena-Herrmann B. Investigating sources of variability in metabolomic data in the EPIC study: the Principal Component Partial R-square (PC-PR2) method. *Metabolomics*, 2014. 40

Ferguson T. S. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2): 209–230, 1973. 25, 87

Finak G., Bashashati A., Brinkman R., and Gottardo R. Merging mixture components for cell population identification in flow cytometry. *Advances in bioinformatics*, 2009: 247646, 2009. 90, 108

Finak G., Perez J.-M., Weng A., and Gottardo R. Optimizing transformations for automated, high throughput analysis of flow cytometry data. *BMC Bioinformatics*, 11: 546, 2010. 91

Flamar A.-L., Xue Y., Zurawski S. M., Montes M., King B., Sloan L., Oh S., Banchereau J., Lévy Y., and Zurawski G. Targeting concatenated HIV antigens to human CD40 expands a broad repertoire of multifunctional CD4+ and CD8+ T cells. *AIDS (London, England)*, 27 (April): 2041–51, 2013. 32

Flynn B. J., Kastenmüller K., Wille-Reece U., Tomaras G. D., Alam M., Lindsay R. W., Salazar A. M., Perdiguero B., Gomez C. E., Wagner R., Esteban M., Park C. G., Trumpfheller C., Keler T., Pantaleo G., Steinman R. M., and Seder R. Immunization with HIV Gag targeted to dendritic cells followed by recombinant New York vaccinia virus induces robust T-cell immunity in nonhuman primates. *Proceedings of the National Academy of Sciences of the United States of America*, 108(17): 7131–7136, 2011. 32

Foulkes A. *Applied statistical genetics with R: For Population-based Association studies*. Use R! Springer Verlag, 2009. 122

Fraley C. and Raftery A. E. Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering. *Journal of Classification*, 24(2): 155–181, 2007. 168

Frühwirth-Schnatter S. and Pyne S. Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics*, 11(2): 317–36, 2010. 15, 16, 25, 91, 92, 94, 101, 160

Furman D., Jojic V., Kidd B., Shen-Orr S., Price J., Jarrell J., Tse T., Huang H., Lund P., Maecker H. T., Utz P. J., Dekker C. L., Koller D., and Davis M. M. Apoptosis and other immune biomarkers predict influenza vaccine responsiveness. *Molecular systems biology*, 9 (659): 659, 2013. 23, 24, 73, 74, 105, 141

Gaucher D., Therrien R., Kettaf N., Angermann B. R., Boucher G., Filali-Mouhim A., Moser J. M., Mehta R. S., Drake D. R., Castro E., Akondy R., Rinfret A., Yassine-Diab B., Said E. a., Chouikh Y., Cameron M. J., Clum R., Kelvin D., Somogyi R., Greller L. D., Balderas R. S., Wilkinson P., Pantaleo G., Tartaglia J., Haddad E. K., and Sékaly R.-P. Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses. *The Journal of experimental medicine*, 205(13): 3119–3131, 2008. 23, 73

Ge Y. and Sealfon S. C. flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding. *Bioinformatics*, 28(15): 2052–8, 2012. 26, 101, 102, 107

Gelman A., Carlin J. B., Stern H. S., Dunson D. B., Vehtari A., and Rubin D. B. *Bayesian Data Analysis*. Texts in Statistical Science. Chapman and Hall/CRC, 3 edition, 2013. 87

Geman S. and Geman D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6: 721 – 741, 1984. 87

Genuer R., Poggi J.-M., and Tuleau-Malot C. Variable selection using random forests. *Pattern Recognition Letters*, 31(14): 2225–2236, 2010. 33, 67

Gilks W. R., Richardson S., and Spiegelhalter D. J. *Markov Chain Monte Carlo in practice*. Interdisciplinary Statistics. Chapman and Hall, 1 edition, 1996. 87

Goeman J. J. and Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8): 980–7, 2007. 18, 46

Granich R. M., Gilks C. F., Dye C., De Cock K. M., and Williams B. G. Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model. *Lancet*, 373(9657): 48–57, 2009. 32

Grant R. M., Lama J. R., Anderson P. L., McMahan V., Liu A. Y., Vargas L., Goicochea P., Casapía M., Guanira-Carranza J. V., Ramirez-Cardich M. E., Montoya-Herrera O., Fernández T., Veloso V. G., Buchbinder S. P., Chariyalertsak S., Schechter M., Bekker L.-G., Mayer K. H., Kallás E. G., Amico K. R., Mulligan K., Bushman L. R., Hance R. J., Ganoza C., Defechereux P., Postle B., Wang F., McConnell J. J., Zheng J.-H., Lee J., Rooney J. F., Jaffe H. S., Martinez A. I., Burns D. N., and Glidden D. V. Preexposure chemoprophylaxis for HIV prevention in men who have sex with men. *The New England journal of medicine*, 363(27): 2587–99, 2010. 32

Grossman C. Interactions between the gonadal steroids and the immune system. *Science*, 227 (4684): 257–261, 1985. 73

Guedj M., Robin S., Celisse A., and Nuel G. Kerfdr: a semi-parametric kernel-based approach to local false discovery rate estimation. *BMC Bioinformatics*, 10: 84, 2009. 33

Guo X., Qi H., Verfaillie C. M., and Pan W. Statistical significance analysis of longitudinal gene expression data. *Bioinformatics*, 19(13): 1628–1635, 2003. 44

Hartemink A. J., Gifford D. K., Jaakoola T. S., and Young R. A. Maximum-likelihood estimation of optimal scaling factors for expression array normalization. In *BiOS 2001 The International Symposium on Biomedical Optics*, pages 132–140. International Society for Optics and Photonics, 2001. 39

Harville D. A. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, 72(358): 320–338, 1977. 51

Hastie T. Generalized additive models. In Chambers J. M. and Hastie T. J., editors, *Statistical Models in S*, chapter 7. Wadsworth & Brooks/Cole, 1992. 48

Hitchcock D. B., Booth J. G., and Casella G. The effect of pre-smoothing functional data on cluster analysis. *Journal of Statistical Computation and Simulation*, 77(12): 1043–1055, 2007. 51

Hoerl A. and Kennard R. Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*, 12: 55–67, 1970. 33

Hu Y., Gao L., Shi K., and Chiu D. K. Y. Detection of deregulated modules using deregulatory linked path. *PloS one*, 8(7): e70412, 2013. 44

Huang A. and Wand M. P. Simple Marginally Noninformative Prior Distributions for Covariance Matrices. *Bayesian Analysis*, 8(2): 439–452, 2013. 94

Hummel M., Meister R., and Mansmann U. GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics*, 24(1): 78–85, 2008. 27, 41, 44, 46, 58, 60, 61, 105

Illumina I. Whole-Genome Gene Expression Direct Hybridization Assay Guide, 2010. 38

Irizarry R. A., Hobbs B., Collin F., Beazer-Barclay Y. D., Antonellis K. J., Scherf U., and Speed T. P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2): 249–264, 2003. 39

Jirtle R. L. and Skinner M. K. Environmental epigenomics and disease susceptibility. *Nature reviews Genetics*, 8(4): 253–262, 2007. 71

Johnson W. E., Li C., and Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1): 118–27, 2007. 20, 40, 52

Juárez M. A. and Steel M. F. J. Model-Based Clustering of Non-Gaussian Panel Data Based on Skew- t Distributions. *Journal of Business & Economic Statistics*, 28(1): 52–66, 2010. 94

Kalli M., Griffin J. E., and Walker S. G. Slice sampling mixture models. *Statistics and Computing*, 21(1): 93–105, 2011. 26, 95

Kanehisa M. and Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1): 27–30, 2000. 16, 33, 42, 43, 61, 125

Kessler D. C., Hoff P. D., and Dunson D. B. Marginally specified priors for non-parametric bayesian estimation. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 77(1): 35–58, 2015. 95

Kim S.-Y. and Volsky D. J. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6: 144, 2005. 41

Klein S. L. and Pekosz A. Sex-based biology and the rational design of influenza vaccination strategies. *The Journal of infectious diseases*, 209(S3): S114–9, 2014. 34

Klein S. L. and Poland G. a. Personalized vaccinology: one size and dose might not fit both sexes. *Vaccine*, 31(23): 2599–2600, 2013. 23, 73

Klein S. L., Jedlicka A., and Pekosz A. The Xs and Y of immune responses to viral vaccines. *The Lancet infectious diseases*, 10(5): 338–349, 2010. 23, 34, 73

Klein S. The effects of hormones on sex differences in infection: from genes to behavior. *Neuroscience and Biobehavioral Reviews*, 24(6): 627–638, 2000. 23, 73

Knight J. C. Genomic modulators of the immune response. *Trends in genetics*, 29(2): 74–83, 2013. 71

Laird N. and Ware J. Random-effects models for longitudinal data. *Biometrics*, 38(4): 963–974, 1982. 46

Lau J. W. and Green P. J. Bayesian Model-Based Clustering Procedures. *Journal of Computational and Graphical Statistics*, 16(3): 526–558, 2007. 97

Law C. W., Chen Y., Shi W., and Smyth G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, 15(2): R29, 2014. 27, 106

Le Cao K.-A. and Le Gall C. Integration and variable selection of ' omics ' data sets with PLS : a survey. *Journal de la Société Française de Statistique*, 152(2): 77–96, 2011. 131, 136, 139

Le Cao K.-A., Rossouw D., Robert-Granié C., and Besse P. A sparse PLS for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology*, 7(1): Article 35, 2008. 22, 131, 139

Le Cao K.-A., Martin P. G. P., Robert-Granié C., and Besse P. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, 10 (34), 2009. 131

114

Leek J. T. and Storey J. D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genetics*, 3(9): 12, 2007. 40

Leek J. T., Scharpf R. B., Bravo H. C., Simcha D., Langmead B., Johnson W. E., Geman D., Baggerly K., and Irizarry R. a. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature reviews Genetics*, 11(10): 733–9, 2010. 40

Lévy Y., Thiébaut R., Montes M., Lacabaratz C., Sloan L., King B., Pérusat S., Harrod C., Cobb A., Roberts L. K., Surenaud M., Boucherie C., Zurawski S., Delaugerre C., Richert L., Chêne G., Banchereau J., and Palucka K. Dendritic cell-based therapeutic vaccine elicits polyfunctional HIV-specific T-cell immunity associated with control of viral load. *European journal of immunology*, 44(9): 2802–10, 2014. 22, 23, 43, 52, 53, 54, 65, 66

Lewden C., Chêne G., Morlat P., Raffi F., Dupon M., Dellamonica P., Pellegrin J.-L., Katlama C., Dabis F., Leport C., CO3 t. A. C. A.-C., ANRS, and Groups A. S. HIV-infected adults with a CD4 cell count greater than 500 cells/mm3 on long-term combination antiretroviral therapy reach same mortality rates as the general population. *Journal of Acquired Immune Deficiency Syndromes*, 46(1): 72–77, 2007. 31

Li J., Bushel P. R., Chu T.-M., and Wolfinger R. D. Principal Variance Components Analysis: Estimating Batch Effects in Microarray Gene Expression Data. In Scherer A., editor, *Batch Effects and Noise in Microarray Experiment*, chapter 12. John Wiley & Sons, Ltd, Chichester, UK, 2009. 40

Li S., Rouphael N., Duraisingham S., Romero-Steiner S., Presnell S., Davis C., Schmidt D. S., Johnson S. E., Milton A., Rajam G., Kasturi S., Carlone G. M., Quinn C., Chaussabel D., Palucka a. K., Mulligan M. J., Ahmed R., Stephens D. S., Nakaya H. I., and Pulendran B. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nature immunology*, 15(2), 2013. 42

Lin L., Chan C., Hadrup S. R., Froesig T. M., Wang Q., and West M. Hierarchical Bayesian mixture modelling for antigen-specific T-cell subtyping in combinatorially encoded flow cytometry studies. *Statistical applications in genetics and molecular biology*, 12(3): 309–331, 2013. 90

Liquet B., Le Cao K.-A., Hocini H., and Thiébaut R. A novel approach for biomarker selection and the integration of repeated measures experiments from two assays. *BMC Bioinformatics*, 13(1): 325, 2012. 18, 43

Liu J., Hughes-Oliver J. M., and Menius J. A. Domain-enhanced analysis of microarray data using GO annotations. *Bioinformatics*, 23(10): 1225–34, 2007. 41, 43

Liva S. M. and Voskuhl R. R. Testosterone Acts Directly on CD4+ T Lymphocytes to Increase IL-10 Production. *The Journal of Immunology*, 167(4): 2060–2067, 2001. 73

Lo A. Y. On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 12(1): 351–357, 1984. 25, 89

Lo K., Brinkman R. R., and Gottardo R. Automated gating of flow cytometry data via robust model-based clustering. *Cytometry. Part A : the journal of the International Society for Analytical Cytology*, 73(4): 321–32, 2008. 90

Luan Y. and Li H. Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics*, 20(3): 332–9, 2004. 44

Maciejewski H. Gene set analysis methods: statistical models and methodological differences. *Briefings in bioinformatics*, 15(4): 504–18, 2014. 43

Marbach D., Costello J. C., Küffner R., Vega N. M., Prill R. J., Camacho D. M., Allison K. R., Kellis M., Collins J. J., and Stolovitzky G. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8): 796–804, 2012. 106

Marx V. Biology: The big challenges of big data. *Nature*, 498(7453): 255–60, 2013. 29

McCullagh P. and Nelder J. A. *Generalized linear models.* Chapman and Hall, 1989. 74, 75

Mi H., Muruganujan A., and Thomas P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research*, 41(Database issue): D377–86, 2013. 42

Molenberghs G. and Verbeke G. Likelihood Ratio, Score, and Wald Tests in a Constrained Parameter Space. *The American Statistician*, 61(1): 22–27, 2007. 19, 49

Mosmann T. R., Naim I., Rebhahn J., Datta S., Cavenaugh J. S., Weaver J. M., and Sharma G. SWIFT-scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, Part 2: Biological evaluation. *Cytometry. Part A : the journal of the International Society for Analytical Cytology*, (3), 2014. 91

Murohashi M., Hinohara K., Kuroda M., Isagawa T., Tsuji S., Kobayashi S., Umezawa K., Tojo a., Aburatani H., and Gotoh N. Gene set enrichment analysis provides insight into novel signalling pathways in breast cancer stem cells. *British journal of cancer*, 102(1): 206–12, 2010. 53

Naim I., Datta S., Rebhahn J., Cavenaugh J. S., Mosmann T. R., and Sharma G. SWIFT-scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, Part 1: Algorithm design. *Cytometry Part A*, pages n/a–n/a, 2014. 90

Neal R. M. Slice sampling. *The Annals of Statistics*, 31(3): 705–767, 2003. 26, 95

Nueda M. J., Sebastián P., Tarazona S., García-García F., Dopazo J., Ferrer A., and Conesa A. Functional assessment of time course microarray data. *BMC Bioinformatics*, 10 Suppl 6: S9, 2009. 44, 46

Nueda M. J., Carbonell J., Medina I., Dopazo J., and Conesa A. Serial Expression Analysis: a web tool for the analysis of serial gene expression data. *Nucleic Acids Research*, 38(Web Server issue): W239–W245, 2010. 46, 61

Obermoser G., Presnell S., Domico K., Xu H., Wang Y., Anguiano E., Thompson-Snipes L., Ranganathan R., Zeitner B., Bjork A., Anderson D., Speake C., Ruchaud E., Skinner J., Alsina L., Sharma M., Dutartre H., Cepika A., Israelsson E., Nguyen P., Nguyen Q.-A., Harrod A. C., Zurawski S. M., Pascual V., Ueno H., Nepom G. T., Quinn C., Blankenship D., Palucka K., Banchereau J., and Chaussabel D. Systems Scale Interactive Exploration Reveals Quantitative and Qualitative Differences in Response to Influenza and Pneumococcal Vaccines. *Immunity*, 38(4): 831–844, 2013. 21, 43, 47, 54, 57, 58

Olsen N. J. and Kovacs W. J. Gonadal steroids and immunity. *Endocrine Review*, 17(4): 369–384, 1996. 73

Palermo R. E., Patterson L. J., Aicher L. D., Korth M. J., Robert-Guroff M., and Katze M. G. Genomic Analysis Reveals Pre-and Postchallenge Differences in a Rhesus Macaque AIDS Vaccine Trial: Insights into Mechanisms of Vaccine Efficacy. *Journal of Virology*, 85(2): 1099–1116, 2011. 43

Park T., Yi S.-G., Lee S., Lee S. Y., Yoo D.-H., Ahn J.-I., and Lee Y.-S. Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics*, 19(6): 694–703, 2003. 44

Parker H. S. and Leek J. T. The practical effect of batch on genomic prediction. *Statistical applications in genetics and molecular biology*, 11(3): Article 10, 2012. 40

Patel A. C. Basic science for the practicing physician: gene expression microarrays. *Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology*, 101(3): 325–332, 2008. 38

Pennell L. M., Galligan C. L., and Fish E. N. Sex affects immunity. *Journal of autoimmunity*, 38(2-3): J282–291, 2012. 73

Pitman J. *Combinatorial Stochastic Processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin/Heidelberg, 2006. 16, 88

Prieto C., Rivas M. J., Sánchez J. M., López-Fidalgo J., and De Las Rivas J. Algorithm to find gene expression profiles of deregulation and identify families of disease-altered genes. *Bioinformatics*, 22(9): 1103–10, 2006. 44

Pulendran B. and Ahmed R. Immunological mechanisms of vaccination. *Nature Immunology*, 131(6): 509–517, 2011. 32

Pyne S., Hu X., Wang K., Rossin E., Lin T.-I., Maier L. M., Baecher-Allan C., McLachlan G. J., Tamayo P., Hafler D. a., De Jager P. L., and Mesirov J. P. Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 106(21): 8519–24, 2009. 90, 91, 101, 108

Qian Y., Wei C., Eun-Hyung Lee F., Campbell J., Halliley J., Lee J. a., Cai J., Kong Y. M., Sadat E., Thomson E., Dunn P., Seegmiller A. C., Karandikar N. J., Tipton C. M., Mosmann T., Sanz I. n., and Scheuermann R. H. Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data. *Cytometry. Part B, Clinical cytometry*, 78 Suppl 1(May): S69–82, 2010. 90

Querec T. D., Akondy R. S., Lee E. K., Cao W., Nakaya H. I., Teuwen D., Pirani A., Gernert K., Deng J., Marzolf B., Kennedy K., Wu H., Soumaya B., Herold O., Miller J., Vencio R. Z., Mulligan M., Aderem A., Ahmed R., and Pulendran B. Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nature Immunology*, 10(1): 116–125, 2008. 33, 43

Rajicic N., Cuschieri J., Finkelstein D. M., Miller-Graziano C. L., Hayden D., Moldawer L. L., Moore E., O'Keefe G., Pelik K., Warren H. S., and Schoenfeld D. a. Identification and interpretation of longitudinal gene expression changes in trauma. *PloS one*, 5(12): e14380, 2010. 44

Rappuoli R. and Aderem A. A 2020 vision for vaccines against HIV, tuberculosis and malaria. *Nature*, 473(7348): 463–469, 2011. 32

Ratmann O., Andrieu C., Wiuf C., and Richardson S. Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 106: 10576–10581, 2009. 27, 106

Rerks-Ngarm S., Pitisuttithum P., Nitayaphan S., Kaewkungwal J., Chiu J., Paris R., Premsri N., Namwat C., de Souza M., Adams E., Benenson M., Gurunathan S., Tartaglia J., McNeil J. G., Francis D. P., Stablein D., Birx D. L., Chunsuttiwat S., Khamboonruang C., Thongcharoen P., Robb M. L., Michael N. L., Kunasol P., and Kim J. H. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *New England Journal of Medicine*, 361(23): 2209–2220, 2009. 32

Robinson D. P., Lorenzo M. E., Jian W., and Klein S. L. Elevated 17β-estradiol protects females from influenza A virus pathogenesis by suppressing inflammatory responses. *PLoS pathogens*, 7(7): e1002149, 2011. 73

Robinson P. N. Deep phenotyping for precision medicine. *Human Mutation*, 33: 777–780, 2012. 31

Roederer M., Keele B. F., Schmidt S. D., Mason R. D., Welles H. C., Fischer W., Labranche C., Foulds K. E., Louder M. K., Yang Z.-Y., Todd J.-P. M., Buzby A. P., Mach L. V., Shen L., Seaton K. E., Ward B. M., Bailer R. T., Gottardo R., Gu W., Ferrari G., Alam S. M., Denny T. N., Montefiori D. C., Tomaras G. D., Korber B. T., Nason M. C., Seder R. a., Koup R. a., Letvin N. L., Rao S. S., Nabel G. J., and Mascola J. R. Immunological and virological mechanisms of vaccine-mediated protection against SIV and HIV. *Nature*, 505 (7484): 502–8, 2014. 32

Sakiani S., Olsen N. J., and Kovacs W. J. Gonadal steroids and humoral immunity. *Nature reviews. Endocrinology*, 9(1): 56–62, 2013. 73

Schafer J. L. and Graham J. W. Missing data: our view of the state of the art. *Psychological methods*, 7(2): 147–177, 2002. 139

Schroeder A., Mueller O., Stocker S., Salowsky R., Leiber M., Gassmann M., Lightfoot S., Menzel W., Granzow M., and Ragg T. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC molecular biology*, 7: 3, 2006. 39

Schulze A. and Downward J. Navigating gene expression using microarrays–a technology review. *Nature cell biology*, 3(8): E190–E195, 2001. 38

Self S. G. and Liang K.-y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82 (398): 605–610, 1987. 19, 49

Sethuraman J. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4: 639–650, 1994. 88

Shahbaba B., Tibshirani R., Shachaf C. M., and Plevritis S. K. Bayesian gene set analysis for identifying significant biological pathways. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 60(4): 541–557, 2011. 27, 41, 44, 62, 105

Shen H. and Huang J. Z. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99: 1015–1034, 2008. 135

Shen-Orr S. S., Tibshirani R., Khatri P., Bodian D. L., Staedtler F., Perry N. M., Hastie T., Sarwal M. M., Davis M. M., and Butte A. J. Cell type-specific gene expression differences in complex tissues. *Nature methods*, 7(4): 287–9, 2010. 28, 108

Shi W., Oshlack A., and Smyth G. K. Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. *Nucleic Acids Research*, 38(22): e204, 2010. 20, 39, 52

Simon N. and Tibshirani R. A Permutation Approach to Testing Interactions in Many Dimensions. *arXiv preprint arXiv:1206.6519*, pages 1–33, 2012. 24, 76

Simonini G., Xu Z., Caputo R., De Libero C., Pagnini I., Pascual V., and Cimaz R. Clinical and transcriptional response to the long-acting interleukin-1 blocker canakinumab in Blau syndrome-related uveitis. *Arthritis and rheumatism*, 65(2): 513–518, 2013. 47

Smyth G. K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular*, 3(1), 2004. 41

Snijders T. A. B. and Bosker R. J. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage, 2nd edition, 2012. 51

Srivastava S. and Chen L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Research*, 38(17): e170, 2010. 106

Stigler S. M. Who Discovered Bayes's Theorem ? *The American Statistician*, 37(4a): 290–296, 1983. 85

Storey J. D., Xiao W., Leek J. T., Tompkins R. G., and Davis R. W. Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 102(36): 12837–12842, 2005. 18, 43, 44

Strachan T. and Read A. Nucleic acid hybridization assays. In *Human Molecular Genetics*, chapter 5. New York: Wiley-Liss, 2nd edition, 1999. 38

Stram D. O. and Lee J. W. Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50: 1171–1177, 1994. 49

Stram D. O. and Lee J. W. Corrections to "Variance components testing in the longitudinal mixed effects model" by D. O. Stram and J. W. Lee; 50, 1171-1177, 1994. *Biometrics*, 51(3): 1196, 1995. 49

Su L. F., Kidd B. a., Han A., Kotzin J. J., and Davis M. M. Virus-specific CD4(+) memory-phenotype T cells are abundant in unexposed adults. *Immunity*, 38(2): 373–383, 2013. 73

Subramanian A., Tamayo P., Mootha V. K., Mukherjee S., Ebert B. L., Gillette M. A., Paulovich A., Pomeroy S. L., Golub T. R., Lander E. S., and Mesirov J. P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43): 15545–15550, 2005. 18, 27, 41, 43, 44, 60, 105

Sugár I. P. and Sealfon S. C. Misty Mountain clustering: application to fast unsupervised flow cytometry gating. *BMC Bioinformatics*, 11(1): 502, 2010. 90

Teh Y. W. Dirichlet Process, 2010. 25, 88, 89

Tenenhaus A., Philippe C., Guillemot V., Le Cao K.-A., Grill J., and Frouin V. Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15(3): 569–83, 2014. 133, 140

The Joint United Nations Programme on HIV/AIDS (UNAIDS). Global report: UNAIDS report on the global AIDS epidemic 2013, 2013. 31

The Joint United Nations Programme on HIV/AIDS (UNAIDS). Fact Sheet: Global AIDS epidemic facts and figures 2014, 2014. 31

Thiébaut R., Hejblum B., and Richert L. L'analyse des «Big Data» en recherche clinique. *Revue d'Épidémiologie et de Santé Publique*, 62(1): 1–4, 2014. 29

Tian L., Greenberg S. A., Kong S. W., Altschuler J., Kohane I. S., and Park P. J. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38): 13544–13549, 2005. 46

Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1): 267–288, 1996. 33, 136

Tibshirani R., Walther G., and Hastie T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2): 411–423, 2001. 20, 51

Tracy R. P. 'Deep phenotyping': characterizing populations in the era of genomics and systems biology. *Current opinion in lipidology*, 19: 151–157, 2008. 31

Tusher V. G., Tibshirani R., and Chu G. 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): 5116–21, 2001. 141

Van Buuren S. and Groothuis-Oudshoorn K. Multivariate Imputation by Chained Equations. *Journal Of Statistical Software*, 45(3): 1–67, 2011. 139

van Dyk D. A. and Jiao X. Metropolis-Hastings within Partially Collapsed Gibbs Samplers. *Journal of Computational and Graphical Statistics*, 2014. 26, 95, 162

van Dyk D. A. and Park T. Partially Collapsed Gibbs Samplers. *Journal of the American Statistical Association*, 103(482): 790–796, 2008. 26, 95

Verbeke G. and Molenberghs G. *Linear mixed models for longitudinal data*. Springer Series in Statistics. Springer, 2000. 51

Vinzi V., Trinchera L., and Amato S. Pls path modeling: from foundations to recent developments and open issues for model assessment and improvement. In Vinzi V., Chin W. W., Henseler J., and Wang H., editors, *Handbook of Partial Least Squares*, pages 47–82. Springer, 2010. 133

Walker S. G. Sampling the Dirichlet Mixture Model with Slices. *Communications in Statistics - Simulation and Computation*, 36(1): 45–54, 2007. 26, 95

Wang L., Chen X., Wolfinger R. D., Franklin J. L., Coffey R. J., and Zhang B. A unified mixed effects model for gene set analysis of time course microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 8(1): 1–18, 2009. 18, 41, 43, 44

Wegelin J. A. A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case. Technical report, Department of Statistics, University of Washington, Seattle, 2000. 131

West M. Hyperparameter estimation in Dirichlet process mixture models. In *ISDS discussion paper series*, pages #92–03. Duke University, 1992. 157, 158

Wold H. Estimation of principal components and related models by iterative least squares. In Krishnaiaah P. R., editor, *Multivariate Analysis*, pages 391–420. New York: Academic Press, 1966. 131

Wolfinger R. D., Gibson G., Wolfinger E. D., Bennett L., Hamadeh H., Bushel P., Afshari C., and Paules R. S. Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of computational biology*, 8(6): 625–37, 2001. 44

Wu D. and Smyth G. K. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, (20): 1–12, 2012. 27, 105

Wu H., Lu T., Xue H., and Liang H. Sparse Additive Ordinary Differential Equations for Dynamic Gene Regulatory Network Modeling. *Journal of the American Statistical Association*, 109(506): 700–716, 2014. 27, 106

Wu Z., Irizarry R. A., Gentleman R., Martinez-Murillo F., and Spencer F. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association*, 99(468): 909–917, 2004. 39

Xie Y., Wang X., and Story M. Statistical methods of background correction for Illumina BeadArray data. *Bioinformatics*, 25(6): 751–7, 2009. 39, 52

Yang Y. H., Dudoit S., Luu P., Lin D. M., Peng V., Ngai J., and Speed T. P. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4): e15, 2002. 39

Yekutieli D. and Benjamini Y. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4): 1165–1188, 2001. 49, 123

Zak D. E. and Aderem A. Systems biology of innate immunity. *Immunological reviews*, 227(1): 264–282, 2009. 32

Zak D. E., Andersen-Nissen E., Peterson E. R., Sato A., Hamilton M. K., Borgerding J., Krishnamurty A. T., Chang J. T., Adams D. J., Hensley T. R., Salter A. I., Morgan C. a., Duerr A. C., De Rosa S. C., Aderem A., and McElrath M. J. Merck Ad5/HIV induces broad innate immune activation that predicts CD8+ T-cell responses but is attenuated by preexisting Ad5 immunity. *Proceedings of the National Academy of Sciences of the United States of America*, 109(50): E3503–E3512, 2012. 47

Zare H., Shooshtari P., Gupta A., and Brinkman R. R. Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinformatics*, 11: 403, 2010. 90

Zhang K., Wang H., Bathke A. C., Harrar S. W., Piepho H.-P., and Deng Y. Gene set analysis for longitudinal gene expression data. *BMC Bioinformatics*, 12(1): 273, 2011. 44, 46

Zou H. and Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301–320, 2005. 33, 141

Appendix A:
Multiple testing correction

A.1 Multiple testing

When several tests are performed simultaneously on the same data, there is an increase of the global type I error. The type I error is the error of a false positive when testing H_0 and the type II error is the error of a false negative when testing H_0, as depicted in table A.1. Their respective probabilities are denoted by α and β, defined as follow:

Definition. $\alpha = \mathbb{P}(H_0$ is rejected $\mid H_0$ is true$)$

Definition. $\beta = \mathbb{P}(H_0$ is not rejected $\mid H_0$ is False$)$

	H_0 **not rejected**	H_0 **rejected**
(H_0) **true**	True Negative (TN) *correct*	False Positive (FP) *type I error*
(H_0) **false**	False Negative (FN) *type II error*	True Positive (TP) *correct*

Table A.1 – Possible results of a test

The power of a test can be expressed as $\mathbb{P}(H_0$ is rejected $\mid H_0$ is False$)= 1 - \beta$ and is linked to the type II error probability.

In a lot of applications, especially in biomedical sciences, one wants to control α when performing a test. When performing multiple tests however, the generalization of the type I error probability is subject to discussion. If the Family Wise Error Rate (FWER) seems the natural extension of the type I error, several other from the quantities defined in table A.2, such as the False Discovery Rate (FDR) can be considered.

In the case of gene expression data, the number of tests performed is very large (several). Controlling the FWER is usually too stringent in exploratory genome-wide studies, and the FDR. FWER is more appropriate in confirmatory studies (in genomics, using quantitative Polymerase Chain Reaction – qPCR– to precisely measure gene expression of a few candidate probes).

Reality	H_0 not rejected	H_0 rejected	Total
H_0 True	U	V	m_0
H_0 False	T	S	m_1
Total	W	R	m

Table A.2 – Cross table of reality and decision regarding m test of null hypothesis H_0

A.2   Family Wise Error Rate

The Family Wise Error Rate (FWER) is defined as the probability that at least one of the m tests considered is wrongly rejecting H_0:

Definition. $FWER = \mathbb{P}(V > 0)$

Several procedures exists to control the FWER in an experiment [Dudoit and Van der Laan, 2008]. Historically, a simple and widely used method is the Bonferroni method. If each single test is controlled at the α^* level, then the FWER is bounded:

$$FWER \leq 1 - (1 - \alpha^*)^m$$

A taylor expansion gives:

$$FWER \leq m\alpha^* \leq \min(m\alpha^*, 1)$$

Therefore, controlling each test at the level:

$$\alpha^* = \frac{\alpha}{m}$$

ensures that $FWER$ is controlled at level α.

A.3   False Discovery Rate

Another quantity of interest for multiple testing correction is the expected false positive rate, called False Discovery Rate (FDR):

Definition. $FDR = \mathbb{E}\left(\dfrac{V}{R}\right)$

Controlling the FDR at a level α is less stringent than controlling the FWER at the same level α since $FDR \leq FWER$ [Dudoit and Van der Laan, 2008], with the limit case being when all H_0 are true ($m_1 = 0$) [Foulkes, 2009]. So any procedure controlling the FWER consequently controls the FDR.

Now several procedures exists to control the FDR in an experiment [Dudoit and Van der Laan, 2008]. Especially, the Benjamini & Hochberg procedure for correcting p-values in order to control this FDR is widely used. Let's consider a vector of m raw p-values:

1. the raw p-values are ordered by increasing order: $p_{(1)} \ldots p_{(m)}$, where $p_{(1)}$ is the lowest p-value obtained among the m tests performed, and $p_{(m)}$ the highest

2. For the level α, one seeks the highest k so that: $p_{(k)} \leq \dfrac{k}{m}\alpha$

3. All the null hypothesis $H_{0;(i)}$ are rejected for $i = 1 \ldots k$

Yekutieli and Benjamini [2001] proposed an extension that accounts for all kind of dependencies (the previous procedure was only robust to specifyc type of dependencies between the tests). The only change compared to the Benjamini-Hochberg procedure is in step 2.:

2. For the level α, one seeks the highest k so that: $p_{(k)} \leq \dfrac{k}{m}\dfrac{\alpha}{c(m)}$

where $c(m) = \displaystyle\sum_{i=1}^{m} \dfrac{1}{i}$

Appendix B:

Hand-picked immune-related subsets of KEGG and Gene Ontology databases

Comparison of TcGSA results on other gene sets databases for the DALIA-1 trial

In addition to utilizing the whole blood Illumina V2 modules from Chaussabel et al. [2008] the DALIA-1 data were analyzed with two other databases: i) a subset of the KEGG [Kanehisa and Goto, 2000] pathway database (table B.1) and ii) a subset of the GO [Ashburner et al., 2000] database. Since the whole blood modules were derived from multiple independent datasets that encompass a wide range of immune related diseases (http://www.interactivefigures.com/dm3/vaccine-paper/faq.gsp for more information), the 260 module gene sets are highly enriched in immune related annotations. This facilitates the interpretation of TcGSA findings in terms of describing immune changes in the blood associated with vaccination and viral rebound following treatment interruption. This explains why the modules are more sensitive in pre-ATI during the vaccination phase of the DALIA-1 trial (Figure B.1). However, the viral rebound is such a cataclysm for the immune system that regardless of the database used, a large part of the gene sets are activated. Indeed, during pre-ATI (vaccination phase of the DALIA-1 trial), 3 out of 75 gene sets were significant in the subset of KEGG, and 0 out of 131 in the subset of GO. During post-ATI, 73 out of 75 gene sets were significant in the subset of KEGG, and 101 out of 131 in the subset of GO. 2 gene sets the subset of KEGG and 20 from the subset of GO were automatically discarded because less than 10 probes or more than 500 probes were observed.

Tables B.1 and B.2 two lists of hand-picked immune-related subsets of KEGG and Gene Ontology databases respectively. The GO subset is transversal, that is to say that each chosen gene set is exclusive of the others (if this exclusivity is true for KEGG and Module by definition of their gene set, GO has a tree structure where numerous gene sets are fully encompassed into others).
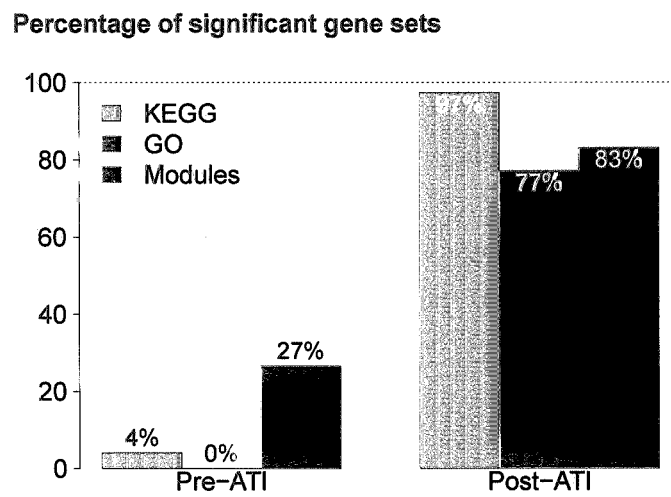
Percentage of significant gene sets

Figure B.1 – Comparison of TcGSA results on DALIA-1 for the three gene sets databases considered

APPENDIX B: KEGG AND GO IMMUNE SUBSETS

	KEGG ID	Description
1	path:hsa04910 †	Endocrine System:Insulin signaling pathway
2	path:hsa04920 †	Endocrine System:Adipocytokine signaling pathway
3	path:hsa03320 †	Endocrine System:PPAR signaling pathway
4	path:hsa04912 †	Endocrine System:GnRH signaling pathway
5	path:hsa04914 †	Endocrine System:Progesterone-mediated oocyte maturation
6	path:hsa04916 †	Endocrine System:Melanogenesis
7	path:hsa04614 †	Endocrine System:Renin-angiotensin system
8	path:hsa04640 †	Immune System:Hematopoietic cell lineage
9	path:hsa04610 †	Immune System:Complement and coagulation cascades
10	path:hsa04620 †	Immune System:Toll-like receptor signaling pathway
11	path:hsa04621 †	Immune System:NOD-like receptor signaling pathway
12	path:hsa04622 †	Immune System:RIG-I-like receptor signaling pathway
13	path:hsa04623 †	Immune System:Cytosolic DNA-sensing pathway
14	path:hsa04650 †	Immune System:Natural killer cell mediated cytotoxicity
15	path:hsa04612 †	Immune System:Antigen processing and presentation
16	path:hsa04660 †	Immune System:T cell receptor signaling pathway
17	path:hsa04662 †	Immune System:B cell receptor signaling pathway
18	path:hsa04664 †	Immune System:Fc epsilon RI signaling pathway
19	path:hsa04666 †	Immune System:Fc gamma R-mediated phagocytosis
20	path:hsa04670 †	Immune System:Leukocyte transendothelial migration
21	path:hsa04672 †	Immune System:Intestinal immune network for IgA production
22	path:hsa04062 †	Immune System:Chemokine signaling pathway
23	path:hsa04510 †	Cell Communication:Focal adhesion
24	path:hsa04520 †	Cell Communication:Adherens junction
25	path:hsa04530 †	Cell Communication:Tight junction
26	path:hsa04540 †	Cell Communication:Gap junction
27	path:hsa04110 †	Cell Growth and Death:Cell cycle
28	path:hsa04114 †	Cell Growth and Death:Oocyte meiosis
29	path:hsa04210 †	Cell Growth and Death:Apoptosis
30	path:hsa04115 †	Cell Growth and Death:p53 signaling pathway
31	path:hsa04144 †	Transport and Catabolism:Endocytosis
32	path:hsa04145 †	Transport and Catabolism:Phagosome
33	path:hsa04142*†	Transport and Catabolism:Lysosome
34	path:hsa04146 †	Transport and Catabolism:Peroxisome
35	path:hsa04140 †	Transport and Catabolism:Regulation of autophagy
36	path:hsa04810 †	Cell Motility:Regulation of actin cytoskeleton
37	path:hsa02010 †	Membrane Transport:ABC transporters
38	path:hsa04010 †	Signal Transduction:MAPK signaling pathway
39	path:hsa04012 †	Signal Transduction:ErbB signaling pathway
40	path:hsa04310 †	Signal Transduction:Wnt signaling pathway
41	path:hsa04330 †	Signal Transduction:Notch signaling pathway
42	path:hsa04340 †	Signal Transduction:Hedgehog signaling pathway
43	path:hsa04350 †	Signal Transduction:TGF-beta signaling pathway
44	path:hsa04370 †	Signal Transduction:VEGF signaling pathway
45	path:hsa04630 †	Signal Transduction:Jak-STAT signaling pathway
46	path:hsa04064 †	Signal Transduction:NF-kappa B signaling pathway
47	path:hsa04020 †	Signal Transduction:Calcium signaling pathway
48	path:hsa04070 †	Signal Transduction:Phosphatidylinositol signaling system
49	path:hsa04151 †	Signal Transduction:PI3K-Akt signaling pathway
50	path:hsa04150 †	Signal Transduction:mTOR signaling pathway
51	path:hsa04080 †	Signaling Molecules and Interaction:Neuroactive ligand-receptor interaction
52	path:hsa04060 †	Signaling Molecules and Interaction:Cytokine-cytokine receptor interaction
53	path:hsa04512 †	Signaling Molecules and Interaction:ECM-receptor interaction
54	path:hsa03030 †	Replication and Repair:DNA replication
55	path:hsa03410 †	Replication and Repair:Base excision repair
56	path:hsa03420 †	Replication and Repair:Nucleotide excision repair
57	path:hsa03430 †	Replication and Repair:Mismatch repair
58	path:hsa03440 †	Replication and Repair:Homologous recombination
59	path:hsa03450	Replication and Repair:Non-homologous end-joining
60	path:hsa03460 †	Replication and Repair:Fanconi anemia pathway
61	path:hsa03060 †	Folding, Sorting and Degradation:Protein export
62	path:hsa04141 †	Folding, Sorting and Degradation:Protein processing in endoplasmic reticulum
63	path:hsa04130 †	Folding, Sorting and Degradation:SNARE interactions in vesicular transport
64	path:hsa04120 †	Folding, Sorting and Degradation:Ubiquitin mediated proteolysis
65	path:hsa04122	Folding, Sorting and Degradation:Sulfur relay system
66	path:hsa03050 †	Folding, Sorting and Degradation:Proteasome
67	path:hsa03018 †	Folding, Sorting and Degradation:RNA degradation
68	path:hsa03010*†	Translation:Ribosome
69	path:hsa00970 †	Translation:Aminoacyl-tRNA biosynthesis
70	path:hsa03013 †	Translation:RNA transport
71	path:hsa03015 †	Translation:mRNA surveillance pathway
72	path:hsa03008 †	Translation:Ribosome biogenesis in eukaryotes
73	path:hsa03020 †	Transcription:RNA polymerase
74	path:hsa03022 †	Transcription:Basal transcription factors
75	path:hsa03040*†	Transcription:Spliceosome

*: significant (FDR<0.05) in pre-ATI   †: significant (FDR<0.05) in post-ATI

Table B.1 – Selected KEGG pathways for investigating DALIA-1

Table B.2 – Selected GO pathways for investigating DALIA-1

	GO ID	Description
1	GO:0002218 †	activation of innate immune response
2	GO:0006956 †	complement activation
3	GO:0002429 †	immune response-activating cell surface receptor signaling pathway
4	GO:0002758 †	innate immune response-activating signal transduction
5	GO:0019883 †	antigen processing and presentation of endogenous antigen
6	GO:0019884 †	antigen processing and presentation of exogenous antigen
7	GO:0048002 †	antigen processing and presentation of peptide antigen
8	GO:0002504 †	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II
9	GO:0002475	antigen processing and presentation via MHC class Ib
10	GO:0002468	dendritic cell antigen processing and presentation
11	GO:0002578	negative regulation of antigen processing and presentation
12	GO:0002579 †	positive regulation of antigen processing and presentation
13	GO:0002577 †	regulation of antigen processing and presentation
14	GO:0002457	T cell antigen processing and presentation
15	GO:0002339	B cell selection
16	GO:0002263 †	cell activation involved in immune response
17	GO:0051607 †	defense response to virus
18	GO:0002432	granuloma formation
19	GO:0002434	immune complex clearance
20	GO:0043299 †	leukocyte degranulation
21	GO:0001909 †	leukocyte mediated cytotoxicity
22	GO:0019724 †	B cell mediated immunity
23	GO:0002228 †	natural killer cell mediated immunity
24	GO:0002707 †	negative regulation of lymphocyte mediated immunity
25	GO:0002708 †	positive regulation of lymphocyte mediated immunity
26	GO:0002706 †	regulation of lymphocyte mediated immunity
27	GO:0002456 †	T cell mediated immunity
28	GO:0002444 †	myeloid leukocyte mediated immunity
29	GO:0002704 †	negative regulation of leukocyte mediated immunity
30	GO:0002705 †	positive regulation of leukocyte mediated immunity
31	GO:0002703 †	regulation of leukocyte mediated immunity
32	GO:0002522	leukocyte migration involved in immune response
33	GO:0002698 †	negative regulation of immune effector process
34	GO:0008228	opsonization
35	GO:0002699 †	positive regulation of immune effector process
36	GO:0002697 †	regulation of immune effector process
37	GO:0002679 †	respiratory burst involved in defense response
38	GO:0002250 †	adaptive immune response
39	GO:0002367 †	cytokine production involved in immune response
40	GO:0006959 †	humoral immune response
41	GO:0002418 †	immune response to tumor cell
42	GO:0002437 †	inflammatory response to antigenic stimulus
43	GO:0006957 †	complement activation, alternative pathway
44	GO:0001867	complement activation, lectin pathway
45	GO:0002227	innate immune response in mucosa
46	GO:0045824 †	negative regulation of innate immune response
47	GO:0045089 †	positive regulation of innate immune response
48	GO:0045088 †	regulation of innate immune response
49	GO:0034341 †	response to interferon-gamma
50	GO:0034340 †	response to type I interferon
51	GO:0050777 †	negative regulation of immune response
52	GO:0002251 †	organ or tissue specific immune response
53	GO:0052555	positive regulation by organism of immune response of other organism involved in symbiotic interaction
54	GO:0002821 †	positive regulation of adaptive immune response
55	GO:0002922	positive regulation of humoral immune response
56	GO:0002839 †	positive regulation of immune response to tumor cell
57	GO:0002863 †	positive regulation of inflammatory response to antigenic stimulus
58	GO:0002830	positive regulation of type 2 immune response
59	GO:0002765 †	immune response-inhibiting signal transduction
60	GO:0002768 †	immune response-regulating cell surface receptor signaling pathway
61	GO:0052552	modulation by organism of immune response of other organism involved in symbiotic interaction
62	GO:0002819 †	regulation of adaptive immune response
63	GO:0002718 †	regulation of cytokine production involved in immune response
64	GO:0043309	regulation of eosinophil degranulation
65	GO:0002920 †	regulation of humoral immune response

66	GO:0002837 †	regulation of immune response to tumor cell
67	GO:0002861 †	regulation of inflammatory response to antigenic stimulus
68	GO:0033006 †	regulation of mast cell activation involved in immune response
69	GO:0043380	regulation of memory T cell differentiation
70	GO:0043313	regulation of neutrophil degranulation
71	GO:0045622 †	regulation of T-helper cell differentiation
72	GO:0002828 †	regulation of type 2 immune response
73	GO:0042092 †	type 2 immune response
74	GO:0002520	immune system development
75	GO:0002366 †	leukocyte activation involved in immune response
76	GO:0050902	leukocyte adhesive activation
77	GO:0042113 †	B cell activation
78	GO:0001767	establishment of lymphocyte polarity
79	GO:0001771	immunological synapse formation
80	GO:0002285 †	lymphocyte activation involved in immune response
81	GO:0030098 †	lymphocyte differentiation
82	GO:0046651 †	lymphocyte proliferation
83	GO:0030101 †	natural killer cell activation
84	GO:0051250 †	negative regulation of lymphocyte activation
85	GO:0031294 †	lymphocyte costimulation
86	GO:0050871 †	positive regulation of B cell activation
87	GO:0045621 †	positive regulation of lymphocyte differentiation
88	GO:0050671 †	positive regulation of lymphocyte proliferation
89	GO:0032816 †	positive regulation of natural killer cell activation
90	GO:0050870 †	positive regulation of T cell activation
91	GO:0050864 †	regulation of B cell activation
92	GO:0045619 †	regulation of lymphocyte differentiation
93	GO:0050670 †	regulation of lymphocyte proliferation
94	GO:0032814 †	regulation of natural killer cell activation
95	GO:0050863 †	regulation of T cell activation
96	GO:0050868 †	negative regulation of T cell activation
97	GO:0046634 †	regulation of alpha-beta T cell activation
98	GO:0046643 †	regulation of gamma-delta T cell activation
99	GO:2001188	regulation of T cell activation via T cell receptor contact with antigen bound to MHC molecule on antigen presenting cell
100	GO:0045580 †	regulation of T cell differentiation
101	GO:0042129 †	regulation of T cell proliferation
102	GO:0046631 †	alpha-beta T cell activation
103	GO:0001768	establishment of T cell polarity
104	GO:0046629 †	gamma-delta T cell activation
105	GO:0035709	memory T cell activation
106	GO:0002286 †	T cell activation involved in immune response
107	GO:0030217 †	T cell differentiation
108	GO:0042098 †	T cell proliferation
109	GO:0002274 †	myeloid leukocyte activation
110	GO:0002695 †	negative regulation of leukocyte activation
111	GO:0002696 †	positive regulation of leukocyte activation
112	GO:0043030 †	regulation of macrophage activation
113	GO:0033003 †	regulation of mast cell activation
114	GO:0030885	regulation of myeloid dendritic cell activation
115	GO:0001776 †	leukocyte homeostasis
116	GO:0050900 †	leukocyte migration
117	GO:0002262 †	myeloid cell homeostasis
118	GO:0002683 †	negative regulation of immune system process
119	GO:0050857 †	positive regulation of antigen receptor-mediated signaling pathway
120	GO:0060369	positive regulation of Fc receptor mediated stimulatory signaling pathway
121	GO:0002253 †	activation of immune response
122	GO:0002687 †	positive regulation of leukocyte migration
123	GO:0070426	positive regulation of nucleotide-binding oligomerization domain containing signaling pathway
124	GO:2000525	positive regulation of T cell costimulation
125	GO:0002645 †	positive regulation of tolerance induction
126	GO:0034123 †	positive regulation of toll-like receptor signaling pathway
127	GO:0002440 †	production of molecular mediator of immune response
128	GO:0002682	regulation of immune system process
129	GO:0002200 †	somatic diversification of immune receptors
130	GO:0045058 †	T cell selection
131	GO:0002507 †	tolerance induction

*: significant (FDR<0.05) in pre-ATI   †: significant (FDR<0.05) in post-ATI

Table B.2: Selected GO pathways for investigating DALIA-1

Appendix C:
Partial Least Squares methods

Partial Least Squares is an exploratory method that was originally proposed in the field of chemometrics [Wold, 1966]. There exists numerous different algorithms implementing the Partial Least Squares [Wegelin, 2000; Boulesteix and Strimmer, 2007]. Indeed, there are several ways of presenting the method, mainly related to the application field. The focus here is on the views presented in the works by Le Cao et al. [2008, 2009]; Le Cao and Le Gall [2011], that are oriented toward genomics.

C.1   Partial Least Squares

C.1.2   Regression approach

Introduction

Let's consider the following linear regression $Y = X\beta + \epsilon$ with:
- n observations
- $Y_{n \times q}$, q variables to be explained à expliquer
- $X_{n \times p}$, p explicative variables
- $\beta_{p \times q}$, regression coefficients
- $\epsilon_{n \times q}$, errors

In the scope of such a regression model, X is assumed of full rank[1] in order to be able to estimate its parameters according to the Ordinary Least Squares method[2] (OLS). Yet this hypothesis is not always true. Indeed, in the high dimensional case where $dim(X) > n$, or yet if some explicative variables are colinear, for instance, then X is not of full rank. And OLS estimate is not unique. Partial Least Squares (PLS) methods are a way to solve such problems.

---

1. $X_{n \times p}$ is said to be of full rank if and only if $rank(X) = p$
2. OLS estimates of β is: $\widehat{\beta} = (X'X)^{-1}X'Y$. Its computation requires therefore $X'X$ to be invertible. If $n << p$ or if there are colinearities in X columns, then $X'X$ is not invertible.

PLS method principle

The PLS method seeks to construct latent variables or scores, orthogonal between each other, which are linear combinations of the original explicative variables[1]. It is therefore important that those original explicative variables are centered to avoid to give more weights to some of them in regards of others. It is also customary to reduce those original explicative variables in order to for their coefficients to be comparable. However, one must be careful because such an operation artificially increases the variability of original explicative variables whose variance was originally very low. This leads to noise amplification. To avoid such an issue, one must carefully pre-select explicative variables (only take into account explicative variables with a significant variability).

PLS1 algorithm

For starters, let's consider that Y contains only one column: $q = 1$. Thanks to the PLS1 algorithm, the following linear model is fitted:

$$Y = T\gamma + \epsilon_H \quad \text{with} \begin{cases} T = \begin{pmatrix} \xi_1 & \cdots & \xi_H \end{pmatrix} \\[2mm] \gamma = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_H \end{pmatrix} \\[2mm] \epsilon_H = P_{(\xi_1,\dots,\xi_H)^\perp} Y \end{cases}$$

where, by design, the ξ_h are orthogonal with one another. The ξ_h are called **latent variables** of X, or X *scores*, h going from 1 to H, where H is the final number of scores kept in the model.

PLS2 algorithm

Let's consider now the case where there are $q > 1$ variables to explain: $Y \in \mathcal{M}_{n \times q}$. PLS method's idea is to seek latent scores $(\xi_1, \dots, \xi_H)$ of X which *best explains* Y, that is to say which maximize covariance between X and Y. Therefore, latent scores of Y are also computed and optimized:

$$(\xi_h, \omega_h) \quad = \quad \underset{\substack{\xi \in \{X_h u, \|u\|=1\} \\ \omega \in \{Y_h v, \|v\|=1\}}}{\arg\max} \; <\xi, \omega> \quad = \quad \underset{\substack{\xi \in \{X_h u, \|u\|=1\} \\ \omega \in \{Y_h v, \|v\|=1\}}}{\arg\max} \; cov(\xi, \omega)$$

The equation to solve is thus: $\underset{\|u\|=1, \; \|v\|=1}{\arg\max} \; cov(X_h u, Y_h v)$.

---

1. X columns

PLS1 Algorithm

---

1. Step 1

$$X_1 = X \text{ and } Y_1 = Y$$

$$\xi_1 = \underset{\xi \in \{X_1 u, \|u\|=1\}}{\arg\max} \ < \xi, Y_1 >$$

$$Y_1 = \xi_1 \gamma_1 + \epsilon_1$$

where γ_1 is estimated by OLS

and $\epsilon_1 = P_{\xi_1^\perp} Y_1$ (orthogonal projection)

2. Step 2

$$X_2 = P_{\xi_1^\perp} X_1 \text{ and } Y_2 = \epsilon_1$$

$$\xi_2 = \underset{\xi \in \{X_2 u, \|u\|=1\}}{\arg\max} \ < \xi, Y_2 >$$

$$Y_2 = \xi_2 \gamma_2 + \epsilon_2 \qquad \text{where } \epsilon_2 = P_{\xi_2^\perp} Y_2$$

$$\vdots$$

3. Step h $(h \leq dim(X))$

$$X_h = P_{\xi_{h-1}^\perp} X_{h-1} \text{ and } Y_h = \epsilon_{h-1}$$

$$\xi_h = \underset{\xi \in \{X_h u, \|u\|=1\}}{\arg\max} \ < \xi, Y_h >$$

$$Y_h = \xi_h \gamma_h + \epsilon_h \qquad \text{where } \epsilon_h = P_{\xi_h^\perp} Y_h$$

$$\vdots$$

H steps

We respectively name u_h and v_h the X and Y **loadings**. They are defined by $\xi_h = X_h u_h$ and $\omega_h = Y_h v_h$, where $(\omega_1, \ldots, \omega_H)$ are the latent scores from Y.

Remark: In the regression perspective, only the ξ_h are of interest. The ω_h being not orthogonal to one another, their interpretation is dubious and they are useful only in the derivation of the ξ_h.

C.1.3 Canonical perspective

This can also referred be referred a symmetrical perspective [Tenenhaus et al., 2014] or "mode A" [Vinzi et al., 2010]. In this case, there are two groums of variables, one group in X and the other one in Y. There is no prior knowledge about the direction of any potential association between variables from these two groups. As before, couples of latent scores whose covariance is maximum are constructed from both X and Y:

$(\xi_1, \omega_1), \ldots, (\xi_H, \omega_H)$:

$$(\xi_h, \omega_h) = \underset{\substack{\xi \in \{X_h u, \|u\|=1\} \\ \omega \in \{Y_h v, \|v\|=1\}}}{\arg\max} \; <\xi, \omega> \; = \underset{\substack{\xi \in \{X_h u, \|u\|=1\} \\ \omega \in \{Y_h v, \|v\|=1\}}}{\arg\max} \; cov\,(\xi, \omega)$$

The only difference resides in the deflation of the Y matrix in the iterative algorithm. It is not deflated on the X scores (as in the regression perspective), but on the Y scores. This has the effect of deriving successive ω_h that are orthogonal to one another this time.

Results exploitation One can then perform a *communality analysis*, which focus on the correlations between the original data (X and Y) and their latent scores (($\xi_1, \ldots, \xi_H$) et ($\omega_1, \ldots, \omega_H$)). One can also be interested in the relations between X and Y, for example by representing their projections in the latent scores subspaces.

C.1.4 General PLS algorithm

The general PLS algorithm with $q \geq 1$ can thus be written as follows:

PLS agorithm

---

1. $X_0 = X$ and $Y_0 = Y$

2. For $h = 1 \ldots H$

   (a) solve: $(u_h, v_h) = \underset{\|u\|=1, \; \|v\|=1}{\arg\max} \; < X_{h-1}u, Y_{h-1}v >$

   (b) $\xi_h = X_{h-1}u_h$

   $\omega_h = Y_{h-1}v_h$

   (c) $X_h = X_{h-1} - \xi_h \left(X'_{h-1}\dfrac{\xi_h}{\| \xi_h \|^2} \right)'$

   (d) $Y_h = \begin{cases} Y_{h-1} - \xi_h \left(Y'_{h-1}\dfrac{\xi_h}{\| \xi_h \|^2} \right)' & regression \\[2em] Y_{h-1} - \omega_h \left(Y'_{h-1}\dfrac{\omega_h}{\| \omega_h \|^2} \right)' & canonic \end{cases}$

C.2 The sparse Partial Least Squares method

C.2.1 sPLS algorithm

The sparse PLS (*sPLS*) method objective is to get sparse loadings, that is to say loadings with very few non zero elements, in order to ease the component interpretation.

Matrix Singular Value Decompostion (SVD)

Definition: $\forall \ M \in \mathcal{M}_{p \times q}$ of rank r, $\exists \ U \in \mathcal{M}_{p \times r}$, $\Delta \in \mathcal{D}_{r \times r}$, $V \in \mathcal{M}_{q \times r}$ so that

$$M = U \Delta V'$$

- Δ is diagonal, its element δ_i are the singular value of M
- $\delta_i \geq 0$, $\forall i = 1 \ldots r$
- U's columns are orthogonal to one another
- V's columns are orthogonal to one another

Property: M SVD decomposition is unique if the δ_i are ordered in a decreasing way and if U and V's columns are normed.

The column vectors of U and V are respectively the left and right singular vectors of M.

Property: If $M = X'Y$ and $M = U \Delta V'$ its SVD decomposition, then the column vectors of U and V are respectively the loadings of X and Y in the PLS model.

Indeed $(u_1, \delta_1 v_1) = \underset{u,v}{\arg \min} \ \| \ M - uv' \ \|_F^2$, where $\| \ M \ \|_F^2 = tr \left(MM' \right)$ (the Frobenius norm), and $(u_2, \delta_2 v_2) = \underset{u,v}{\arg \min} \ \| \ (M - u_1 \delta_1 v_1') - uv' \ \|_F^2$. It follows that $\sum_{i=1}^k \delta_i u_i v_i'$ is the best M approximation of rank k in the sense of Frobenius norm $\| \cdot \|_F$ [Shen and Huang, 2008]. Hence:

$$\underset{\|u\|=1, \ \|v\|=1}{\arg \max} \ < Xu, Yv > \ = \ \underset{\|u\|=1, \ \|v\|=1}{\arg \min} \ \| \ X'Y - \delta_1 uv' \ \|_F^2$$

where δ_1 is the first singular value from $M = X'Y$. This results allow to rewrite the covariance maximization constraint in the PLS and sPLS algorithms [Shen and Huang, 2008].

Penalization

In order to perform sPLS objective (getting sparse loadings), penalties are added to the maximization equation. Those penalties enforce the following idea: the more non

zero coordinates there are on the loadings, the higher is the corresponding penalty. The LASSO [Tibshirani, 1996] constraint (L_1) fits this description, and has the advantage of giving true zero coordinates on the loadings (not just very low coordinates as one can get with classic PLS algorithms). Such a penalty is written: $P_\lambda(t) = \sum_{i=1}^{H} 2\lambda|t_i|$ (where $t = (t_1, \ldots, t_H)$ and λ is a threshold parameter)[1]. In practice, an approximate solution to this LASSO constraint can be computed by using the *soft-thresholding* function [Le Cao and Le Gall, 2011] (this approximation is exact in the case of independent predictors as is the case in the *canonic* perspective).

sPLS algorithm

$P_{\lambda_1}(u)$ and $P_{\lambda_2}(v)$ penalization functions are defined as previously, constraining u_h and v_h computation. The sPLS algorithm can be written as follows:

sPLS algorithm

1. $X_0 = X$ and $Y_0 = Y$

2. For $h = 1 \ldots H$

   (a) $M_{h-1} = X'_{h-1} Y_{h-1}$

   (b) solve $(u_h, v_h) = \underset{\|u\|=1, \|v\|=1}{\arg\min} \| M_{h-1} - \delta_1^{(h)} uv' \|_F^2 + P_{\lambda_1}(u) + P_{\lambda_2}(v)$

   (c) $\xi_h = X_{h-1} u_h$

   $\omega_h = Y_{h-1} v_h$

   (d) $X_h = X_{h-1} - \xi_h \left(X'_{h-1} \dfrac{\xi_h}{\| \xi_h \|^2} \right)'$

   (e) $Y_h = \begin{cases} Y_{h-1} - \xi_h \left(Y'_{h-1} \dfrac{\xi_h}{\| \xi_h \|^2} \right)' & regression \\[2em] Y_{h-1} - \omega_h \left(Y'_{h-1} \dfrac{\omega_h}{\| \omega_h \|^2} \right)' & canonic \end{cases}$

<u>*Remark*</u>: If $P_{\lambda_1}(u) = P_{\lambda_2}(v) = 0$, then one gets the standard PLS algorithm.

---

1. because $|\sum_{i=1}^{H} t_i| \leq \sum_{i=1}^{H} |t_i|$

Minimization : detail of step 2.(b) of the sPLS algorithm

Let's consider the following function to be minimized:

$$\| M_{h-1} - \delta_1^{(h)} uv' \|_F^2 + P_{\lambda_1}(u) + P_{\lambda_2}(v) \tag{35}$$

The LASSO penalty can be approached by the soft-thresholding (with equivalence reached in the case of orthogonal predictors). The minimization problem above (35) can thus be solved by using the soft-thresholding until convergence is reached for $\widehat{u_h}$ and $\widehat{v_h}$:

$$\begin{cases} \widehat{u_h}^{(k+1)} = \dfrac{g_\lambda(M_{h-1}\widehat{v_h}^{(k)})}{\| g_\lambda(M_{h-1}\widehat{v_h}^{(k)}) \|} \\[2em] \widehat{v_h}^{(k+1)} = \dfrac{g_\lambda(M'_{h-1}\widehat{u_h}^{(k)})}{\| g_\lambda(M'_{h-1}\widehat{u_h}^{(k)}) \|} \end{cases}$$

with $g_\lambda(y) = sign(y) \cdot (|y| - \lambda) \cdot \mathbb{1}_{\{(|y|-\lambda)>0\}}$ the soft-thresholding function for each y coordinate of the vectors. Indeed, for a fixed u in (35), one seeks $\widehat{v}$ that minimize:

$$\| M_{h-1} - \delta_1^{(h)} uv' \|_F^2 + P_{\lambda_2}(v) = \sum_{i=1}^{p}\sum_{j=1}^{q}(m_{ij} - u_i v_j)^2 + \sum_{j=1}^{q} P_{\lambda_2}(v_j) \tag{36}$$

with $m_{ij} = (M_{h-1})_{ij}$

Since $\sum_{i=1}^{p} u_i^2 = 1$ (normed u), expanding the squares yields:

$$\sum_{i=1}^{p}(m_{ij} - u_i v_j)^2 = \sum_{i=1}^{p} m_{ij}^2 - 2\sum_{i=1}^{p} m_{ij}u_i v_j + \sum_{i=1}^{p} u_i^2 v_j^2$$

$$= \sum_{i=1}^{p} m_{ij}^2 - 2(M'_{h-1}u)_j v_j + v_j^2$$

and (36) then becomes:

$$\sum_{j=1}^{q}\left(\sum_{i=1}^{p} m_{ij}^2 - 2(M'_{h-1}u)_j v_j + v_j^2 + P_{\lambda_2}(v_j)\right) \tag{37}$$

Thus, for a given u,

$$\widehat{v} = \underset{\|v\|=1}{\arg\min}\ \sum_{j=1}^{q}\left[-2(M'_{h-1}u)_j v_j + v_j^2 + P_{\lambda_2}(v_j)\right]$$

(37) can be minimized for each coordinate j of $\widehat{v}$ independently, $\widehat{v}$ being standardized once each of its coordinate is estimated, and then:

$$\widehat{v_j} = \arg\min\ v_j^2 - 2(M'_{h-1}u)_j v_j + P_{\lambda_2}(v_j) \tag{38}$$

What's more, by derivation of (38) one gets:

$$\frac{\mathrm{d}}{\mathrm{d}v}\left[v_j^2 - 2(M'_{h-1}u)_j v_j + P_{\lambda_2}(v_j)\right] = 0$$

$$\Leftrightarrow \quad v_j = (M'_{h-1}u)_j + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}v}\left[P_{\lambda_2}(v_j)\right] \tag{39}$$

Let's write $z = (M'_{h-1}u)_j$, $y = v_j$ and $\lambda = \lambda_2$. Then (39) becomes:

$$y = z + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}y}\left[P_\lambda(y)\right] \tag{40}$$

If $P_\lambda(v) = 2\lambda|v|$, then :

$$
\begin{aligned}
(40) \Leftrightarrow v &= (z - \lambda) & &\cdot \mathbb{1}_{\{v>0\}} & &+ (z + \lambda) & &\cdot \mathbb{1}_{\{v<0\}} \\
&= (z - \lambda) & &\cdot \mathbb{1}_{\{z-\lambda>0\}} & &+ (z + \lambda) & &\cdot \mathbb{1}_{\{z+\lambda<0\}} \\
&= sign(z)(|z| - \lambda) & &\cdot \mathbb{1}_{\{z>\lambda\}} & &+ sign(z)(|z| - \lambda) & &\cdot \mathbb{1}_{\{z<\lambda\}} \\
&= sign(z)(|z| - \lambda) & &\cdot \mathbb{1}_{\{|z|>\lambda\}} \\
&= g_\lambda(z)
\end{aligned}
$$

Finally, this applies also to u when v is fixed.

By initializing $\widehat{u_h}^{(0)}$ and $\widehat{v_h}^{(0)}$ with the first couple of singular vectors of M_{h-1} SVD decomposition (which is the solution for $\lambda = 0$), convergence is rapidly reached towards the sparse optimum.

The difference between canonical and regression perspectives

The difference between those 2 perspectives takes place during Y update, that is to say at the steps (d) and (e) respectively of the PLS and sPLS algorithms.

- In the regression perspective, the Y matrix is deflated in regards of the X latent scores subspace (by projecting Y onto the space orthogonal to the X latent scores):

$$Y_h \leftarrow Y_{h-1} - \xi_h \left(Y'_{h-1}\frac{\xi_h}{\parallel \xi_h \parallel^2}\right)'$$

- In the canonical perspective, the Y matrix is deflated in regards of its own latent scores subspace (by projecting Y onto the space orthogonal to the Y latent scores)):

$$Y_h \leftarrow Y_{h-1} - \omega_h \left(Y'_{h-1}\frac{\omega_h}{\parallel \omega_h \parallel^2}\right)'$$

- There exist a third perspective, the invariante perspective, where Y is never modified during the algorithm: $Y_h \leftarrow Y_{h-1}$.

Missing values

A simple way of dealing with missing values from either X or Y in PLS methods is to replace those by 0 during the computation of the latent scores (since X has to be centered and reduced, it is desirable that Y be too). But it seems more reasonable to treat missing values before applying any PLS methods, for instance by using a multiple imputation method [Schafer and Graham, 2002; Van Buuren and Groothuis-Oudshoorn, 2011].

C.2.2 Penalization parameters tuning

λ_1 and λ_2 are the two penalty parameter, enforcing the sparsity level on X and Y respectively. There are various way of tuning those parameters [Le Cao et al., 2008]. A common approach is to use cross validation to optimize a given criterion of model quality, simultaneously for both λ_1 and λ_2, one component at a time. For instance, one can use a sequential cross-validated (preferably leave-one-out) optimization of the mean R^2 between the Y variables selected by a given component explained by the corresponding X component on a grid of explored values for λ_1 and λ_2. Alternatively, the Mean Squared Error of Prediction (MSEP) can be used instead of the R^2. Often, instead of tuning λ_1 and λ_2, one tunes the number of variables selected (which is equivalent with the soft-thresholding approximation of the LASSO solution). In practice, those penalization criteria can also be tuned empirically depending on the application problem, so that enough information is conveyed by each component for further interpretation (for instance enrichment annotation of selected genes in genomics applications) [Le Cao and Le Gall, 2011].

C.2.3 Choice of the number of components

In a PLS (or sPLS) model, the maximum number of components possible to derive is p if a regression perspective is adopted, and $min(q, p)$ if a canonical perspective is adopted. But often only a few number of components are analysed.

Let's define:

$$Q_h^2 = 1 - \frac{\sum_{j=1}^q PRESS_h^j}{\sum_{j=1}^q RSS_{h-1}^j}$$

with $PRESS_h^j = \sum_{i=1}^n (y_i^j - \hat{y}_{h(-i)}^j)^2$ the PRediction Error Sum of Squares at step $h-1$ and $RSS_h^j = \sum_{i=1}^n (y_i^j - \hat{y}_h^j)^2$ the Residual Sum of Squares at step $h-1$. $\hat{y}_{h(-i)}^j$ refers to the leave-one-out estimator of observation i. This Q^2 criterion can be seen as the marginal contribution of the latent score ξ_h [Le Cao et al., 2008]. A heuristic cutoff is to keep

including the component $h + 1$ as long as:

$$Q^2_{h+1} \geq (1 - 0.95^2) = 0,0975$$

This limit corresponds to $\sqrt{PRESS_{h+1}} < 0.95\sqrt{RSS_h}$.

The Q^2 is only available in the regression perspective, and if Y latent scores are not penalized (indeed, if the variables contributing to the $PRESS$ are not the same as the one selected at $h - 1$ and contributing to the RSS, the Q^2 criterion is not interpretable). In such cases other empirical methods, for instance relying for instance on cross-validated Root Mean Squared Error of Prediction (RMSEP), can be used.

C.3   Related approaches

Tenenhaus et al. [2014] proposed a general framework that encompass the case of the sPLS method, and that also extend it both to the multi-block case and to an optimization criterion for component that can depend on both the covariance and the correlation.

Appendix D:

Supplementary information on the systems analysis of sex differences in the response to influenza vaccination

D.1 Construction of the gene modules used in the flu vaccination system analysis

First, gene probes were filtered by variance and normalized. Then, hierarchical agglomerative clustering was performed to derive 109 modules. Those 109 are thus data driven and based and co-expression.

In addition, for each gene module, a set of regulatory genes (regulatory program) was assigned based on regression analysis of genes in the modules onto expression of transcription factors. This was conducted using the LARS-EN algorithm [Zou and Hastie, 2005]. The LARS-EN algorithm provides fits of increasing numbers of predictors. To select the best model among the outputs of LARS-EN, we assessed the quality of the resulting models by the Akaike information criterion (AIC) [Akaike, 1974], with sample-specific terms weighted by module variance. The fit with the best AIC score was selected for each module. Detailed statistical procedures have been described [Furman et al., 2013].

D.2 Baseline sex differences

To determine the differences in the baseline's immune measures in males versus females, we investigated univariate association of the 278 measured variables with sex, while controlling the FDR (Appendix A page 121) at a 10% level. This association was estimated using the Significant Analysis of Microarrays approach (SAM) from Tusher et al. [2001]. This approach focus on estimating the signal-to-noise ratio between a two conditions for an important number of covariates. It deals with high-dimensional setting

($n = 87$ and $p = 278$ in our case) by investigating univariate associations while controlling the FDR.

Seven variables were found significantly associated with sex whith an FDR < 0.1, six of which were increased in females. Strikingly, these included several known markers of inflammation, such as LEPTIN, interleukin (IL)-1 receptor agonist (RA), C-reactive protein (CRP), Granulocyte macrophage-colony stimulating factor (GM-CSF), and interleukin IL-5, as well as the phosphorylation levels of STAT3 proteins in unstimulated monocytes (M-pSTAT3). The last significant variable was the gene module 106. It was up-regulated in males compared with females (a significant fraction of this gene module is composed of genes located on the Y chromosome).

Appendix E:

Original *PNAS* article systems analysis of sex differences in the response to influenza vaccination

Systems analysis of sex differences reveals an immunosuppressive role for testosterone in the response to influenza vaccination

David Furman[a,1,2,3], Boris P. Hejblum[b,1], Noah Simon[c], Vladimir Jojic[d], Cornelia L. Dekker[e], Rodolphe Thiébaut[b], Robert J. Tibshirani[c,f], and Mark M. Davis[a,g,h,3]

[a]Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA 94305-5323; [b]ISPED-Epidemiologie-Biostatistique and Institut National de la Santé et de la Recherche Médicale (INSERM), Centre INSERM U897, University of Bordeaux, and INRIA–Statistics in System Biology and Translational Medicine Team, F-33000 Bordeaux, France; [c]Department of Statistics, Stanford University, Stanford, CA 94305-4065; [d]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3175; [e]Department of Pediatrics, Division of Infectious Diseases, Stanford University School of Medicine, Stanford, CA 94305-5208; [f]Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA 94305-5405; [g]Institute for Immunity, Transplantation and Infection, Stanford University, Stanford, CA 94305-5124; and [h]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305-5323

Contributed by Mark M. Davis, November 21, 2013 (sent for review September 23, 2013)

Females have generally more robust immune responses than males for reasons that are not well-understood. Here we used a systems analysis to investigate these differences by analyzing the neutralizing antibody response to a trivalent inactivated seasonal influenza vaccine (TIV) and a large number of immune system components, including serum cytokines and chemokines, blood cell subset frequencies, genome-wide gene expression, and cellular responses to diverse in vitro stimuli, in 53 females and 34 males of different ages. We found elevated antibody responses to TIV and expression of inflammatory cytokines in the serum of females compared with males regardless of age. This inflammatory profile correlated with the levels of phosphorylated STAT3 proteins in monocytes but not with the serological response to the vaccine. In contrast, using a machine learning approach, we identified a cluster of genes involved in lipid biosynthesis and previously shown to be up-regulated by testosterone that correlated with poor virus-neutralizing activity in men. Moreover, men with elevated serum testosterone levels and associated gene signatures exhibited the lowest antibody responses to TIV. These results demonstrate a strong association between androgens and genes involved in lipid metabolism, suggesting that these could be important drivers of the differences in immune responses between males and females.

aging | gender | immuno-endocrine | sexual dimorphism | immunosenescence

The variability in the biology of human populations poses significant challenges in understanding different disease outcomes and developing successful therapeutics. The sources of this variation are likely the consequence of genetics, epigenetics, and the history of antigenic exposure (1, 2). As therapies targeting immune function are developed to improve clinical outcomes in cancer, viral and bacterial infections, autoimmune diseases, and transplantation, identifying the sources of immunological variation and finding biomarkers for immune health and dysfunction are crucial for their success (3).

An important source of immunological variation is known to be the sex of the individual. Males experience a greater severity and prevalence of bacterial, viral, fungal, and parasitic infections than females, who also exhibit a more robust response to antigenic challenges such as infection and vaccination (4, 5). This stronger immune response in females could also explain why they more frequently develop immune-mediated pathologies during influenza infection, such as an overproduction of cytokines (cytokine storm) that contribute to an increase in capillary permeability and lung failure (6). Furthermore, females are at a higher risk for developing autoimmune diseases. In this later context, it is interesting to note that a recent study showed that females had, on average,

1.7 times the frequency of self-specific T cells as males (7). Despite the fact that initial observations relating the sex of the individual with the immune response were made many years ago (8), little is known about the mechanisms underlying these differences.

Some sex-specific variations in the immune response can be directly attributed to sex hormones (9). In humans, sex steroids can bind to intracellular receptors located in immune cells such as monocytes, B cells, and T cells and activate hormone-responsive genes, suggesting that they can directly affect sex-related differences in both innate and adaptive immune responses (10). Whereas estrogens are associated with inflammation and can stimulate proliferation and differentiation of lymphocytes and monocytes, androgens suppress the activity of immune cells by increasing the synthesis of anti-inflammatory cytokines (11, 12).

To date, no clear associations have been found between biological and clinical differences in the immune response between

---

Significance

There are marked differences between the sexes in their immune response to infections and vaccination, with females often having significantly higher responses. However, the mechanisms underlying these differences are largely not understood. Using a systems immunology approach, we have identified a cluster of genes involved in lipid metabolism and likely modulated by testosterone that correlates with the higher antibody-neutralizing response to influenza vaccination observed in females. Moreover, males with the highest testosterone levels and expression of related gene signatures exhibited the lowest antibody responses to influenza vaccination. This study generates a number of hypotheses on the sex differences observed in the human immune system and their relationship to mechanisms involved in the antibody response to vaccination.

---

males and females in humans. In one study, results from public gene expression data (13) showed that many of the genes induced by a yellow fever vaccine were preferentially activated in females (14). However, whether these differences correlate with poor antibody outcomes remains to be determined.

In this study, we sought to determine whether we could identify biomarkers from peripheral blood that could explain the sex-related differences in the serological response to the trivalent inactivated seasonal influenza vaccine (TIV) in both young and older cohorts.

Young and older females had higher neutralizing antibodies than age-matched males, consistent with previous reports (15). Females also showed higher expression of inflammatory markers. However, none of these specific sex-related differences correlated with the observed disparities in the antibody response to TIV. Nevertheless, using a machine learning approach, we identified a set of genes previously shown to be regulated by testosterone and participating in lipid biosynthesis, whose expression was negatively associated with antibody responses to TIV in the male subjects in our study. Moreover, males with high levels of serum testosterone and expressing related gene signatures in blood cells showed the lowest neutralizing responses to TIV. These results suggest that testosterone might be immunosuppressive in vivo in humans, and indicate that its effect on an influenza vaccine and other immune responses could be due to the regulation of genes implicated in the metabolism of lipids.

Results

Elevated Levels of Neutralizing Antibodies upon Influenza Vaccination and Inflammatory Markers in Serum from Females Versus Males. To study the differences in males' versus females' immune systems, we used data from a vaccination and systems immunology study conducted on 91 individuals (37 males and 54 females) of different ages (20–30 and 60–>89 y old) (Table 1) that we recently reported (16). We studied a variety of immune parameters from peripheral blood before vaccination, including cytokines, chemokines, and growth factors in serum, frequencies of diverse blood cell subsets, phosphorylation levels of signal transducer and activator of transcription (STAT) proteins in multiple cells stimulated with a variety of cytokines or unstimulated (96 conditions in total), and whole-blood gene expression. The gene expression data were reduced to 109 gene modules by cluster analysis and assignment of a set of transcription factors (regulatory program) to each gene module as described (16) (SI Materials and Methods). Four individuals were removed from the analysis: two outliers and two with incomplete datasets.

To determine the magnitude of the antibody response to TIV, we performed virus microneutralization assays. The seroconversion rate (percent of individuals with a fourfold or greater change in their post- versus prevaccination microneutralization titer) was computed for each group and strain in the vaccine (SI Materials and Methods). We conducted logistic regression analysis on each of the titer changes (corresponding to the H1N1, H3N2, and B strains) and included the age and sex variables in the model, because age was expected to modify vaccine responses. Females had a greater response than males to the H3N2 strain ($P = 0.0027$) and to a lesser extent to the B strain ($P = 0.02$). In contrast, despite a strong

Table 1. Subjects' baseline characteristics

	Males	Females	P value
Number of subjects	37	54	—
Age range (median), y	20–>89 (63)	20–>89 (68)	0.14
BMI range (median)	19–36 (25)	18–47 (24)	0.61
Cytomegalovirus (+), %	61	57	0.73
Epstein–Barr virus (+), %	70	54	0.13

BMI, body mass index.

Table 2. Age and sex effects on microneutralization antibody titer responses to influenza vaccination

		Beta	SE	z value	P value
H1N1	(Intercept)	-0.272	0.229	-1.190	0.234
	Age	-0.690	0.236	-2.919	0.004
	Sex	-0.011	0.234	-0.047	0.962
H3N2	(Intercept)	-0.038	0.228	-0.166	0.868
	Age	-0.190	0.236	-0.804	0.421
	Sex	-0.716	0.239	-2.992	0.003
B	(Intercept)	-0.502	0.236	-2.128	0.033
	Age	-0.583	0.246	-2.367	0.018
	Sex	-0.594	0.256	-2.324	0.020

age effect ($P = 0.0035$), no differences according to sex were found for the H1N1 strain (Table 2).

To determine the differences in the baseline's immune measures in males versus females, we conducted differential expression analysis across a total of 278 parameters using significance analysis of microarrays (SAM) (17) and found significant differences in 7 parameters [false discovery rate (FDR) $Q < 0.1$], 6 of which were increased in females (Fig. 1). Strikingly, these included several known markers of inflammation, such as LEPTIN, interleukin (IL)-1 receptor agonist (RA), C-reactive protein (CRP), Granulocyte macrophage-colony stimulating factor (GM-CSF), and Interleukin IL-5, as well as the phosphorylation levels of STAT3 proteins in unstimulated monocytes (M-pSTAT3).

One parameter (gene module 106) was up-regulated in males compared with females. A significant fraction of this gene module is composed of genes located on the Y chromosome (enrichment $P < 10^{-9}$) (Table S1). Interestingly, genes participating in the activation of v-akt murine thymoma viral oncogene homolog (Akt) and phospholipase C (PLC) proteins such as mature T-cell proliferation 1 (MTCP1) and phosphatidylinositol-specific phospholipase C, X domain containing 1 (PLCXD1), respectively, clustered with Y chromosome-linked genes in module 106. The regulatory program derived for module 106 (Fig. S1) included genes previously shown to be differentially regulated in males versus females, such as CLOCK (18), ENY2 (19), and IRF1 and IRF7 (20).

It has long been noted that inflammatory markers, especially IL-6, TNF-α, and IL-1β, among others, are increased in the elderly. Thus, we divided the individuals by age group (young, 20–30 y old; old, 60–>89 y old) and investigated whether these differences were also observed in our aging cohort. The serum levels of LEPTIN, IL1-RA, CRP, GM-CSF, and IL-5 were all higher in females regardless of age group ($P < 0.05$) (Fig. 2). However, the differences for CRP, IL1-RA, and LEPTIN were less pronounced in the older group due to an overall increase in the levels of these proteins in older compared with younger males ($P = 0.007$, by Fisher's combined probability). Strikingly, M-pSTAT3 levels were significantly higher in females among young subjects ($P = 0.002$) but not in the elderly ($P = 0.268$), where both sexes had similar levels to those found in young females (Fig. 2). This suggested that other cytokines that signal through STAT3 (e.g., IL-6, IL-11, and LIF, among others) were elevated in both males and females in the older cohort. Because IL-6 is one of the hallmark cytokines of aging, we directly compared IL-6 levels in young versus older (without correction for multiple comparisons) and noticed elevated levels in elderly subjects (Fig. S2), consistent with multiple previous reports.

To identify associations between these sex-related features, we generated correlation matrices and conducted unsupervised clustering with or without IL-6. Interestingly, M-pSTAT3 clustered with CRP and GM-CSF (Fig. S3A), or with CRP, IL-6, and GM-CSF when IL-6 was incorporated in the clustering analysis (Fig. S3B), suggesting that the intracellular levels of phosphorylated
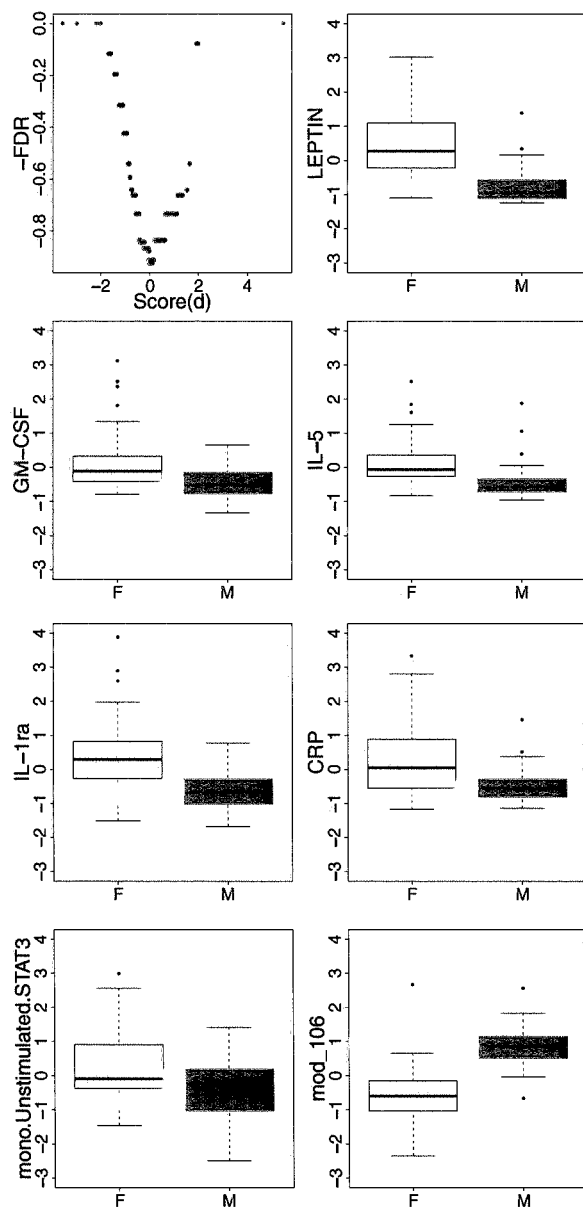
Fig. 1. Significant differences in baseline immune parameters between females and males. Expression of a total of 278 immune features and gene modules was compared between females (F) (*n* = 53) and males (M) (*n* = 34) of different age groups, including serum cytokines, chemokines, and growth factors; frequencies of over 15 blood cell subsets; phosphorylation events in multiple immune cells; and whole-genome gene expression using SAM. A cutoff of *Q* < 0.1 and absolute score(d) > 2 was considered significant (volcano plot; *Top, Left*). Inflammatory markers including LEPTIN, IL-1RA, and CRP and other serum proteins were elevated in females compared with males. mono.Unstimulated.STAT3, baseline levels of pSTAT3 in isolated monocytes. A single gene module (module 106) (*Bottom, Right*) was differentially expressed (and up-regulated in males). Module 106 is enriched for genes located on the Y chromosome ($P < 10^{-9}$). Lower whisker represents the minimum value, lower hinge the first quartile, upper hinge the third quartile, and upper whisker the maximum value. Outliers are represented by circles.

STAT proteins in monocytes likely represent a functional readout corresponding to an inflammatory environment in vivo.

These results indicate that females have a stronger neutralizing response to influenza vaccination and an increased inflammatory serum profile, which correlates with the baseline levels of phosphorylated STAT3 proteins in peripheral monocytes.

Weaker Vaccine Responses in Males with High Expression of Genes Involved in Lipid Metabolism.

To identify features associated with the observed sex differences in vaccine responsiveness, we focused on the neutralizing activity to H3N2 because the largest differences between males and females were found for this strain. Also, this influenza strain is important in a public health context because it causes the highest rates of morbidity and mortality during the influenza season (21). An individual was considered a responder if they had the standard fourfold or greater change in their post- versus prevaccination microneutralization titer (seroconversion). For H3N2, 33 females and 10 males were responders and 20 females and 24 males were nonresponders. We first searched for possible confounding factors by investigating which features could substantially modify the observed sex effect. In brief, we performed forward stepwise logistic regression with sex as the initial predictor. In stepwise regression, each of the immune features was incorporated into the model iteratively and statistics were computed to account for modifications in the regression coefficient of sex (*SI Materials and Methods*). The iterations were stopped when the added feature did not modify the regression coefficient of sex to the standard threshold 20%. By this procedure, we identified two possible confounders: a gene module enriched for genes encoding for ribosomal proteins (module 042) (enrichment $P < 10^{-6}$) and the acute-phase inflammatory marker CRP. Thus, we generated a first model (model 1), which included the variables of sex, module 042, and CRP. The resulting regression coefficient for sex in model 1, after adjusting for confounders, was −2.03 compared with −1.38 (sex variable alone) (Fig. S4).

We then searched for gene expression profiles that could explain the differences in vaccine responsiveness between males and females, namely features having different effects in males or females, while adjusting for confounding variables. To do so, we used the Interact package (*SI Materials and Methods*), which searches for significant interactions between predictors using permutation methods. A significant interaction (FDR *Q* < 0.1) was identified for a module enriched for genes participating in lipid biosynthesis (enrichment $P < 0.001$) (module 052) (Table S1). These genes included LTA4H, encoding for leukotriene A4 hydrolase, which converts leukotriene A4 (LTA4) to active LTB4; MIF (macrophage migration inhibitory factor), which plays a role in the anti-inflammatory effects of glucocorticoids (22); PDSS2 (decaprenyl-diphosphate synthase subunit 2), whose product synthesizes the prenyl side chain of coenzyme Q; and PEX5 (peroxisomal biogenesis factor 5), involved in fatty acid metabolism. The gene regulators derived for module 052 (Fig. S1*B*) included CLOCK (activator) and FBJ murine osteosarcoma viral oncogene homolog (FOS), Jun B proto-oncogene (JUNB), and Jun D proto-oncogene (JUND) (repressors), among others. Interestingly, the CLOCK gene is involved in the regulation of circadian rhythms, as well as in lipid metabolism (23).

We then generated a second model (model 2), which included the variables of sex and module 052, the interaction term (sex × module 052), and the covariates CRP and module 042. The resulting odds ratio (OR) estimate for vaccine response based on the expression of module 052 in model 2 was 0.39 [confidence interval (CI), 0.18–0.84] for males and 2.25 (CI, 1.08–4.67) for females (Fig. 3*A*). This indicates that the probability of being a high responder to TIV decreases significantly with an elevated expression of module 052 in males and with decreased expression of module 052 in females. To determine the extent to which module 052 and its interaction with sex contribute to the
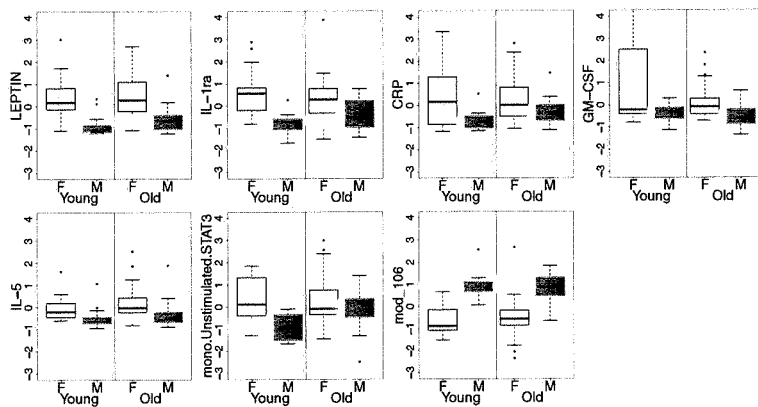
Fig. 2. Differences in baseline immune parameters between females and males by age group. The individuals were first divided by age group (58 older and 29 young), and the significant differences identified between all females (yellow bars) and males (blue bars) using SAM (seven in total; Fig. 1) were used to investigate differences in expression by age group. With the exception of mono.Unstimulated.STAT3, all significant differences between all males and females identified previously were also observed in both age groups ($P < 0.05$). However, the differences in LEPTIN, IL-1RA, and CRP were less pronounced in older individuals due to an overall increase in the levels of these proteins in the serum of older males compared with young males ($P < 0.05$). F, females; M, males. mono.Unstimulated. STAT3, baseline levels of pSTAT3 in isolated monocytes; mod_106, module enriched for Y chromosome genes.

classification model, we computed a cross-validated area under the curve ($_{cv}$AUC) for model 1 and model 2. The $_{cv}$AUC was 0.712 for model 1, and 0.761 for model 2. Furthermore, direct comparison of the two models shows that model 1 is significantly better than model 2 ($P = 0.0019$, by likelihood ratio test).

These results suggest that the observed sex differences in the neutralizing antibody responses to vaccination could be mediated by the expression of genes involved in lipid metabolism.

Blunted Vaccine Response in Males with High Levels of Testosterone and Elevated Expression of Genes Involved in Lipid Metabolism. Our results showing that augmented expression of module 052 correlated with weaker vaccine responsiveness in males but not in females suggested that sex hormones could be involved in expression of this gene module. Indeed, results from chemical–gene interaction analysis (http://ctdbase.org) (24) show that expression of a significant fraction of genes in module 052 can be modulated by testosterone ($P < 0.005$, by hypergeometric test). Thus, we measured free (unbound, bioactive form) testosterone in the sera from the individuals in our study with the hypothesis that, in males, the observed effect of module 052 on vaccine response was dependent on the circulating levels of testosterone. We stratified the male subjects into testosterone high (Thi) or low (Tlo), if they were above or below the median for all of the male subjects (4.06 pg/mL; range, 0.58–24.78 pg/mL), and generated a third model (model 3) for vaccine-neutralizing response, in which male subjects were replaced by Thi ($n = 17$) and Tlo ($n = 17$). The median testosterone level in Thi subjects was 9.55 pg/mL (range, 4.25–24.78 pg/mL), and 2.34 pg/mL (range, 0.58–3.89 pg/mL) in Tlo subjects. The median age for Tlo and Thi males was 77 and 24 y, respectively. Thus, model 3 included the interaction terms module 052 × Thi and module 052 × Tlo, and was also adjusted for age, because of the effect of aging on testosterone levels. Strikingly, the interaction between testosterone levels and module 052 was significant only for the Thi group ($P < 0.005$) and not for Tlo males ($P = 0.18$), and the corresponding OR estimates for vaccine response, according to module 052, were 0.87 (CI, 0.28–2.69) for Tlo and 0.19 (CI, 0.04–0.80) for Thi males. We also tested testosterone levels as a continuous measure by replacing the interaction terms module 052 × Thi and module 052 × Tlo with module 052 × testosterone; the model was also adjusted by sex and age. Consistent with model 3, the interaction of module 052 and testosterone levels was significant ($P = 0.012$) (Fig. S5). This indicates that module 052 has a significant effect on vaccine response in males with high levels of testosterone but not in those with lower levels.

Together, these results show that in males with higher levels of testosterone and elevated expression of genes that participate in lipid metabolism, the antibody response to vaccination is severely

down-regulated, whereas in those with low levels of testosterone, or in females, the contribution of module 052 is not detrimental and the responses to the vaccine remain intact.

Discussion

In this study, we have used a systems approach to the analysis of sex differences in the immune system in humans. These data reinforce and extend previous reports, and point toward a mechanistic hypothesis that may drive the sex disparities observed in responses to vaccination. Differences in vaccine responsiveness in males versus females have been reported for most commercially available vaccines including yellow fever, influenza, measles, mumps, rubella, and hepatitis, among others (5). As in these studies, we find stronger responses to influenza vaccination and significantly increased serum levels of proinflammatory molecules in females compared with males, specifically LEPTIN (25), IL-RA (26), and CRP (27). In addition, we find differences in GM-CSF and IL-5 and in the baseline pSTAT3 levels in monocytes, which correlate with serum CRP, IL-6, and GM-CSF. Consistent with this, LEPTIN and IL-6 activate STAT3 in monocytes, which results in the secretion of CRP and IL1-RA (28, 29). We also find that these sex differences in monocyte pSTAT3 are observed only in young
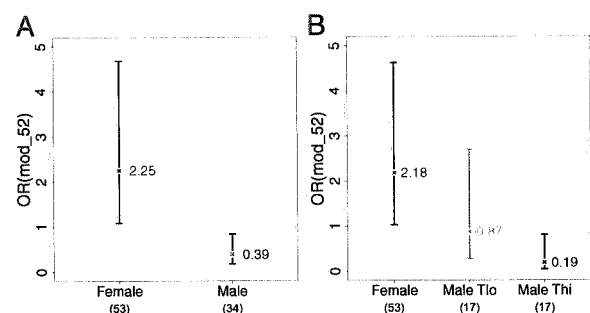

Fig. 3. Odds ratio for vaccine responses in males and females based on expression of module 052. Interaction analysis was conducted for sex and gene expression modules on the serological (microneutralization) responses to TIV (seroconversion to the H3N2 strain). A significant interaction was found between the variables sex and gene module 052. (A) Odds ratio for vaccine response given module 052 in females (red line) and males (blue line). (B) No significant interaction between sex and module 052 is observed for males with low levels of testosterone [Tlo ($n = 17$), brown line], although a significantly negative effect of module 052 is observed for males with high levels of testosterone [Thi ($n = 17$), blue line] (adjusted for confounders including age).

147

subjects, possibly due to increased levels of other cytokines signaling through STAT3 in both sexes from the older cohort (e.g., IL-6). Therefore, the level of pSTAT3 in monocytes likely reflects the sum of diverse inflammatory stimuli targeting STAT3. Consistent with these observations, the phosphorylation levels of STAT3 and other STAT proteins have been found to be increased in asthma (30) and in other inflammatory conditions (31).

With respect to the influenza vaccine response, our results indicate that the natural variation in circulating free testosterone could drive many of the differences observed in the response to vaccines. In particular, males with elevated levels of serum testosterone and high expression of genes participating in lipid metabolism were significantly less likely to respond to TIV. These results are in agreement with previous findings showing an immunosuppressive role of testosterone in animals and in vitro (11, 32) with an increase in the synthesis of anti-inflammatory cytokines such as IL-10 (12). Consistently, men with androgen deficiencies have higher levels of inflammatory cytokines than healthy controls (33). However, we did not find an association between the proinflammatory cytokines that are differentially expressed in females and males and the response to vaccination. Rather, our data indicate that other molecules, such as those involved in lipid biosynthesis, are likely affected by testosterone and modulate the antibody response. In particular, our results suggest that testosterone could act by decreasing expression of transcription factors such as FOS, JUNB, and JUND that, in turn, repress the expression of gene module 052 (Fig. S1B). Consistent with this hypothesis, androgen receptor signaling antagonizes NF-κB and represses AP-1 (FOS/JUN), which mediates the production of proinflammatory and antiviral cytokines (34, 35).

A particularly interesting gene found in module 052 is LTA4H, one of the members of the epoxide hydrolase family. The product of this gene catalyzes the conversion of LTA4 (originating from arachidonic acid) to LTB4, a lipid mediator that has both proinflammatory (via surface receptors) and anti-inflammatory (via activation of peroxisome proliferator-activated receptors and decreased NF-κB expression) activities. Furthermore, LTB4 seems to participate in the differentiation of suppressor cells both from the myeloid (36) as well as from the lymphoid (37) compartments. More generally, several studies in both humans and mice have shown the ability of LTB4 precursors, such as omega-3 and -6 fatty acids, to suppress inflammatory responses (38, 39).

Other potentially relevant genes in module 052 are MIF, PDSS2, and PEX5. MIF participates in the synthesis of prostaglandin E2, a lipid compound that originates from arachidonic acid, binds to immune cells (T cells and dendritic cells), and suppresses inflammatory cytokine production (40, 41). A less clear association is observed for PDSS2, an enzyme that mediates isoprenoid biosynthesis and the incorporation of lipids in proteins. Last, PEX5 participates in the biogenesis of peroxisomes, which regulate various metabolic activities including the degradation of very long chain fatty acids. Peroxisomes have also been implicated in the innate immune system, with a significant reduction observed in inflammation, apparently related to the suppressive effect of TNF-α on peroxisome function (42).

Recent studies have also focused on the possibility that genes located on the X and Y chromosomes also affect the response to vaccination. Polymorphisms in genes on the X chromosome that encode for immunological proteins can influence immune responses to vaccines. For example, Toll-like receptor 7, located on the X chromosome, can escape X inactivation, resulting in higher expression in females than in males (43). Y chromosome genes have also been shown to affect sex-dependent susceptibility to autoimmune disease and possibly to other immune functions (44). However, our results indicate that the expression of genes on the Y chromosome might not be involved in the immune response to vaccines, because the sex-related gene module 106 was not associated with the differences in the response to TIV despite the observation that genes regulating

important immune functions such as the activation of Akt and PLC proteins by MTCP1 and PLCXD1, respectively, are clustered together with Y chromosome genes in this module.

In conclusion, our results are consistent with a large body of work in animals showing that testosterone is immunosuppressive in vivo and extend this to humans responding to a seasonal influenza vaccine and exhibiting typical variations in testosterone levels. We suggest that testosterone acts directly on immune cells by repressing transcription factors (such as FOS, JUN, and others) implicated in immune activation; these transcription factors would in turn repress the expression of genes involved in lipid metabolism with immunosuppressive activities, creating a negative feedback loop.

From an evolutionary perspective, the immunosuppressive effects of testosterone could be advantageous as a possible homeostatic mechanism to turn off the immune response. For instance, experiments with highly pathogenic viruses reconstructed from isolates from the 1918 influenza pandemic (which killed over 50 million people) show that infection with this strain in animal models results in an uncontrolled, deadly cytokine storm (45). Furthermore, suppression of this inflammatory response in infected mice ameliorates immunopathology and decreases mortality (6, 46). It has also been noted that testosterone treatment of castrated male mice made them less susceptible to LPS-induced shock (32). Because males of many species are more likely to experience trauma than females, this positive effect of testosterone may also help to balance out the consequences of reduced immunity to infection.

In summary, we have identified unique proinflammatory markers that are differentially expressed in females compared with males, as well as genes that participate in lipid metabolism that could be modulated by the levels of free testosterone in normal populations and correlate with the sex-related bias in the responsiveness to influenza vaccination.

Materials and Methods

Subjects, Specimens, and Vaccination Protocol. With the exception of the neutralizing antibody response to the vaccine and the determination of testosterone measurements from serum, this study used baseline-level data from a previously published work conducted in 91 healthy donors who were enrolled in an influenza vaccine study at the Stanford-Lucile Packard Children's Hospital (LPCH) Vaccine Program during the 2008–2009 influenza season (16). Thus, only a brief description of the methods is included here. The protocol of the study was approved by the Institutional Review Board of the Research Compliance Office at Stanford University. Blood samples were obtained prevaccination and 28 ± 7 d after receiving a single dose of TIV Fluzone (Sanofi Pasteur). Whole blood was used for gene expression analysis as described (16). Peripheral blood mononuclear cells (PBMCs) were obtained by density gradient centrifugation (Ficoll-Paque) and frozen at –80 °C before transferring to liquid nitrogen. Serum was separated by centrifugation of clotted blood and stored at –80 °C before use. Whole blood, PBMCs, or serum from the first visit (baseline, day 0) was processed and used for determination of gene expression, leukocyte subset frequency, signaling responses to stimulation, serum cytokine and chemokine levels, testosterone levels, and CMV and EBV serostatus by ELISA (Calbiotech). Serum samples from day 0 and day ~28 were used for virus microneutralization titer determination.

Virus Microneutralization Assay. A standard plaque reduction virus microneutralization assay was performed. In brief, serum samples were heat-inactivated for 30 min at 56 °C, serially diluted (twofold) in virus diluent (DMEM, 1% BSA, antibiotics, and 25 mM Hepes), and mixed with 100 median tissue culture infective doses each of the H1N1, H3N2, and B strains (kind gift of George Kemble, MedImmune). Plates were incubated 1 h at 37 °C and 5% CO_2, and 1.5 × 10^4 exponentially growing Madin–Darby canine kidney-London cells (kind gift of David Lewis, Stanford University) were added in 100 μL of virus diluent. Cell cultures were then incubated overnight at 37 °C and 5% CO_2, and washed and fixed with ice-cold acetone for 10 min at room temperature. Fixative was discarded and plates were air-dried. After several washes with washing buffer (PBS, 0.1% Tween-20), wells were incubated for 1 h at room temperature with an anti-influenza A or B nucleoprotein mouse monoclonal antibody (KPL) at 1:4,000 in blocking buffer (PBS, 1% BSA, 0.1% Tween-20). After washing, a secondary antibody (goat anti-mouse IgG, HP-conjugated; KPL) was added at 1:2,000 in blocking buffer and

incubated 1 h at room temperature before revealing with HRP substrate. Absorbance (OD) was read at 490 nm.

Whole-Blood Microarray Analysis of Gene Expression. The procedures for RNA extraction, quantification, hybridization, and scanning were described previously (16). The original microarray probe-level data files can be accessed at the Gene Expression Omnibus repository under accession number GSE41080.

Leukocyte Subset Frequency Determination. PBMCs were thawed, washed with FACS buffer (PBS supplemented with 2% FBS and 0.1% Na azide), and stained with three separate anti-human antibody mixtures containing (*i*) anti-CD3 AmCyan, CD4 Pacific Blue, CD8 allophycocyanin (APC) H7, and CD28 APC; (*ii*) CD3 AmCyan, CD4 Pacific Blue, CD8 APCH7, CD27 PE, and CD45RA PE-Cy5; and (*iii*) CD3 AmCyan, CD19 Alexa Fluor 700, CD56 PE, CD33 PE-Cy7, and TCR APC, all reagents from BD Biosciences. After incubation, cells were washed several times and data were collected using DIVA software on an LSR II instrument (BD Biosciences) and analyzed using FlowJo 8.8.6 (Tree Star).

Phosphorylation of Intracellular Proteins by Phosphoflow. Cells were thawed with FACS buffer and stimulated with the indicated cytokines for 15 min in warm media (RPMI with 10% FBS). Cells were washed and fixed with paraformaldehyde and permeabilized with 95% ice-cold methanol. Different stimulus conditions were bar-coded using a 3 × 3 matrix with Pacific Orange and Alexa Fluor 750 (Invitrogen). Cell mixtures were stained with an antibody mixture as described

previously (16). Data were collected using DIVA software and analyzed with FlowJo 8.8.6.

Serum Cytokine-Level Determination. Cytokines were measured on a Luminex system. Fifty-plex kits were purchased from Millipore and used according to the manufacturer's recommendations with some modifications.

Testosterone-Level Determination. Serum levels of free testosterone were measured using the Free Testosterone ELISA Kit (Calbiotech) as recommended by the manufacturer.

Statistical Analysis. Statistical procedures for gene module construction, interaction analysis, and modeling of vaccine responsiveness can be found in *SI Materials and Methods*.

1. Jirtle RL, Skinner MK (2007) Environmental epigenomics and disease susceptibility. *Nat Rev Genet* 8(4):253–262.
2. Knight JC (2013) Genomic modulators of the immune response. *Trends Genet* 29(2):74–83.
3. Davis MM (2008) A prescription for human immunology. *Immunity* 29(6):835–838.
4. Klein SL (2000) The effects of hormones on sex differences in infection: From genes to behavior. *Neurosci Biobehav Rev* 24(6):627–638.
5. Klein SL, Poland GA (2013) Personalized vaccinology: One size and dose might not fit both sexes. *Vaccine* 31(23):2599–2600.
6. Robinson DP, Lorenzo ME, Jian W, Klein SL (2011) Elevated 17β-estradiol protects females from influenza A virus pathogenesis by suppressing inflammatory responses. *PLoS Pathog* 7(7):e1002149.
7. Su LF, Kidd BA, Han A, Kotzin JJ, Davis MM (2013) Virus-specific CD4(+) memory-phenotype T cells are abundant in unexposed adults. *Immunity* 38(2):373–383.
8. Grossman CJ (1985) Interactions between the gonadal steroids and the immune system. *Science* 227(4684):257–261.
9. Sakiani S, Olsen NJ, Kovacs WJ (2013) Gonadal steroids and humoral immunity. *Nat Rev Endocrinol* 9(1):56–62.
10. Pennell LM, Galligan CL, Fish EN (2012) Sex affects immunity. *J Autoimmun* 38(2-3):J282–J291.
11. Olsen NJ, Kovacs WJ (1996) Gonadal steroids and immunity. *Endocr Rev* 17(4):369–384.
12. Liva SM, Voskuhl RR (2001) Testosterone acts directly on CD4+ T lymphocytes to increase IL-10 production. *J Immunol* 167(4):2060–2067.
13. Gaucher D, et al. (2008) Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses. *J Exp Med* 205(13):3119–3131.
14. Klein SL, Jedlicka A, Pekosz A (2010) The Xs and Y of immune responses to viral vaccines. *Lancet Infect Dis* 10(5):338–349.
15. Cook IF (2008) Sexual dimorphism of humoral immunity with human vaccines. *Vaccine* 26(29-30):3551–3555.
16. Furman D, et al. (2013) Apoptosis and other immune biomarkers predict influenza vaccine responsiveness. *Mol Syst Biol* 9:659.
17. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 98(9):5116–5121.
18. Gómez-Abellán P, et al. (2012) Sexual dimorphism in clock genes expression in human adipose tissue. *Obes Surg* 22(1):105–112.
19. Xiao R, et al. (2012) In utero exposure to second-hand smoke aggravates adult responses to irritants: Adult second-hand smoke. *Am J Respir Cell Mol Biol* 47(6):843–851.
20. Haslinger C, et al. (2004) Microarray gene expression profiling of B-cell chronic lymphocytic leukemia subgroups defined by genomic aberrations and VH mutation status. *J Clin Oncol* 22(19):3937–3949.
21. Thompson WW, et al. (2003) Mortality associated with influenza and respiratory syncytial virus in the United States. *JAMA* 289(2):179–186.
22. Al-Abed Y, et al. (2011) Thyroxine is a potential endogenous antagonist of macrophage migration inhibitory factor (MIF) activity. *Proc Natl Acad Sci USA* 108(20):8224–8227.
23. Turek FW, et al. (2005) Obesity and metabolic syndrome in circadian Clock mutant mice. *Science* 308(5724):1043–1045.
24. Davis AP, et al. (2013) The Comparative Toxicogenomics Database: Update 2013. *Nucleic Acids Res* 41(Database issue):D1104–D1114.
25. Hickey MS, et al. (1996) Gender differences in serum leptin levels in humans. *Biochem Mol Med* 59(1):1–6.
26. Lynch EA, Dinarello CA, Cannon JG (1994) Gender differences in IL-1 alpha, IL-1 beta, and IL-1 receptor antagonist secretion from mononuclear cells and urinary excretion. *J Immunol* 153(1):300–306.
27. Lakoski SG, et al. (2006) Gender and C-reactive protein: Data from the Multiethnic Study of Atherosclerosis (MESA) cohort. *Am Heart J* 152(3):593–598.
28. Zhang D, Sun M, Samols D, Kushner I (1996) STAT3 participates in transcriptional activation of the C-reactive protein gene by interleukin-6. *J Biol Chem* 271(16):9503–9509.
29. Gabay C, Dreyer M, Pellegrinelli N, Chicheportiche R, Meier CA (2001) Leptin directly induces the secretion of interleukin 1 receptor antagonist in human monocytes. *J Clin Endocrinol Metab* 86(2):783–791.
30. Yang XO, et al. (2013) The signaling suppressor CIS controls proallergic T cell development and allergic airway inflammation. *Nat Immunol* 14(7):732–740.
31. O'Shea JJ, Holland SM, Staudt LM (2013) JAKs and STATs in immunity, immunodeficiency, and cancer. *N Engl J Med* 368(2):161–170.
32. Rettew JA, Huet-Hudson YM, Marriott I (2008) Testosterone reduces macrophage expression in the mouse of Toll-like receptor 4, a trigger for inflammation and innate immunity. *Biol Reprod* 78(3):432–437.
33. Malkin CJ, et al. (2004) The effect of testosterone replacement on endogenous inflammatory cytokines and lipid profiles in hypogonadal men. *J Clin Endocrinol Metab* 89(7):3313–3318.
34. Kallio PJ, Poukka H, Moilanen A, Jänne OA, Palvimo JJ (1995) Androgen receptor-mediated transcriptional regulation in the absence of direct interaction with a specific DNA element. *Mol Endocrinol* 9(8):1017–1028.
35. McKay LI, Cidlowski JA (1999) Molecular control of immune/inflammatory responses: Interactions between nuclear factor-kappa B and steroid receptor-signaling pathways. *Endocr Rev* 20(4):435–459.
36. Yokota Y, et al. (2012) Absence of LTB4/BLT1 axis facilitates generation of mouse GM-CSF-induced long-lasting antitumor immunologic memory by enhancing innate and adaptive immune systems. *Blood* 120(17):3444–3454.
37. Juzan M, Hostein N (1992) Role of thymus-eicosanoids in the immune response. *Prostaglandins Leukot Essent Fatty Acids* 46(4):247–255.
38. Simopoulos AP (2002) Omega-3 fatty acids in inflammation and autoimmune diseases. *J Am Coll Nutr* 21(6):495–505.
39. Kanneganti TD, Dixit VD (2012) Immunological complications of obesity. *Nat Immunol* 13(8):707–712.
40. Vassiliou E, Jing H, Ganea D (2003) Prostaglandin E2 inhibits TNF production in murine bone marrow-derived dendritic cells. *Cell Immunol* 223(2):120–132.
41. Jing H, Vassiliou E, Ganea D (2003) Prostaglandin E2 inhibits production of the inflammatory chemokines CCL3 and CCL4 in dendritic cells. *J Leukoc Biol* 74(5):868–879.
42. Schrader M, Fahimi HD (2006) Peroxisomes and oxidative stress. *Biochim Biophys Acta* 1763(12):1755–1766.
43. Pisitkun P, et al. (2006) Autoreactive B cell responses to RNA-related antigens due to TLR7 gene duplication. *Science* 312(5780):1669–1672.
44. Spach KM, et al. (2009) Cutting edge: The Y chromosome controls the age-dependent experimental allergic encephalomyelitis sexual dimorphism in SJL/J mice. *J Immunol* 182(4):1789–1793.
45. Cillóniz C, et al. (2009) Lethal influenza virus infection in macaques is associated with early dysregulation of inflammatory related genes. *PLoS Pathog* 5(10):e1000604.
46. Teijaro JR, et al. (2011) Endothelial cells are central orchestrators of cytokine amplification during influenza virus infection. *Cell* 146(6):980–991.

149

Supporting Information

Furman et al. 10.1073/pnas.1321060111

SI Materials and Methods

Serological Response to Trivalent Inactivated Seasonal Influenza Vaccine. The seroconversion rate (percent of individuals with a fourfold or greater change in their post- versus prevaccination microneutralization titer) was computed for each strain and for each group of individuals. The largest differences between males and females were observed for the H3N2 strain. Thus, the vaccine response was modeled as a binary variable (fold increase ≥4 to the H3N2 strain) in logistic regression analyses.

Gene Module Construction. First, gene probes were filtered by variance and normalized. Hierarchical agglomerative clustering was performed to derive 109 modules. For each gene module, a set of regulatory genes (regulatory program) was assigned based on regression analysis of genes in the modules onto expression of transcription factors. This was conducted using the LARS-EN algorithm (1). The LARS-EN algorithm provides fits of increasing numbers of predictors. To select the best model among the outputs of LARS-EN, we assessed the quality of the resulting models by the Akaike information criterion (AIC) (2), with sample-specific terms weighted by module variance. The fit with the best AIC score was selected for each module. Detailed statistical procedures have been described (3).

Interaction Analysis and Modeling of Vaccine Responsiveness. Potential confounders were identified if they modified the estimates of the sex effect on the vaccine response by more than 20%. A forward strategy was performed starting with a basic model including the sex covariate only.

To identify possible gene module candidates that explain the differences observed in vaccine response, we tested for marginal interactions between gene modules and the sex variable that associate with the neutralization antibody titer outcome (above). To do so, we used the Interact package (http://cran.r-project.org/web/packages/Interact/index.html), which searches for interactions in a binary response model using permutation methods to estimate false discovery rates (FDRs). The significance threshold was set at an FDR of <10% ($Q < 0.1$).

For the estimation of the regression coefficients and odds ratios in the response to vaccination, we conducted simple logistic regression with the categorical variable corresponding to the seroconversion to the H3N2 strain. The following formulas were used in the different models.

Model 1.

$$logit(y_i) = \mu + \beta_s male_i + \beta_c crp_i + \beta_{m42} mod\_42_i + \varepsilon_i$$

Model 2.

$$logit(y_i) = \mu + \beta_s male_i + \beta_c crp_i + \beta_{m42} mod\_42_i + \beta_{m52} mod\_52_i$$
$$+ \beta_{s:m52} male : mod\_52_i + \varepsilon_i$$

Model 3.

$$logit(y_i) = \mu + \beta_c crp_i + \beta_{m42} mod\_42_i + \beta_{m52} mod\_52_i + \beta_a age_i$$
$$+ \beta_{Tlo} maleTlo_i + \beta_{Thi} maleThi_i + \beta_{Tlo:m52} maleTlo:mod\_52_i$$
$$+ \beta_{Thi:m52} maleThi : mod\_52_i + \varepsilon_i,$$

where y_i is the binary response to H3N2 for the ith individual, μ is the average response for females, $male_i$ is a dichotomic variable (1 for male, 0 for female), crp_i is the CRP level of the ith individual, mod_42_i is the 42nd module median expression level of the ith individual, mod_52_i is the 52nd module expression level of the ith individual, $maleTlo_i$ is a dichotomic variable (1 for males with low levels of testosterone, that is, below the median of the male group; 0 for females and males with high levels of testosterone; see below), $maleTlo_i:mod\_52_i$ is the interaction term of mod_52_i and $maleTlo_i$ for the ith individual, $maleThi_i$ is a dichotomic variable (1 for males with high levels of testosterone, that is, above the median of the male group; 0 for females and for males with low levels of testosterone), $maleThi_i:mod\_52_i$ is the interaction term of mod_52_i and $maleThi_i$, and ε_i is the error term for the ith individual.

1. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* 67:301–320.
2. Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Automat Contr* 19(6):716–723.

3. Furman D, et al. (2013) Apoptosis and other immune biomarkers predict influenza vaccine responsiveness. *Mol Syst Biol* 9:659.

150

A   Regulatory Program Module 106
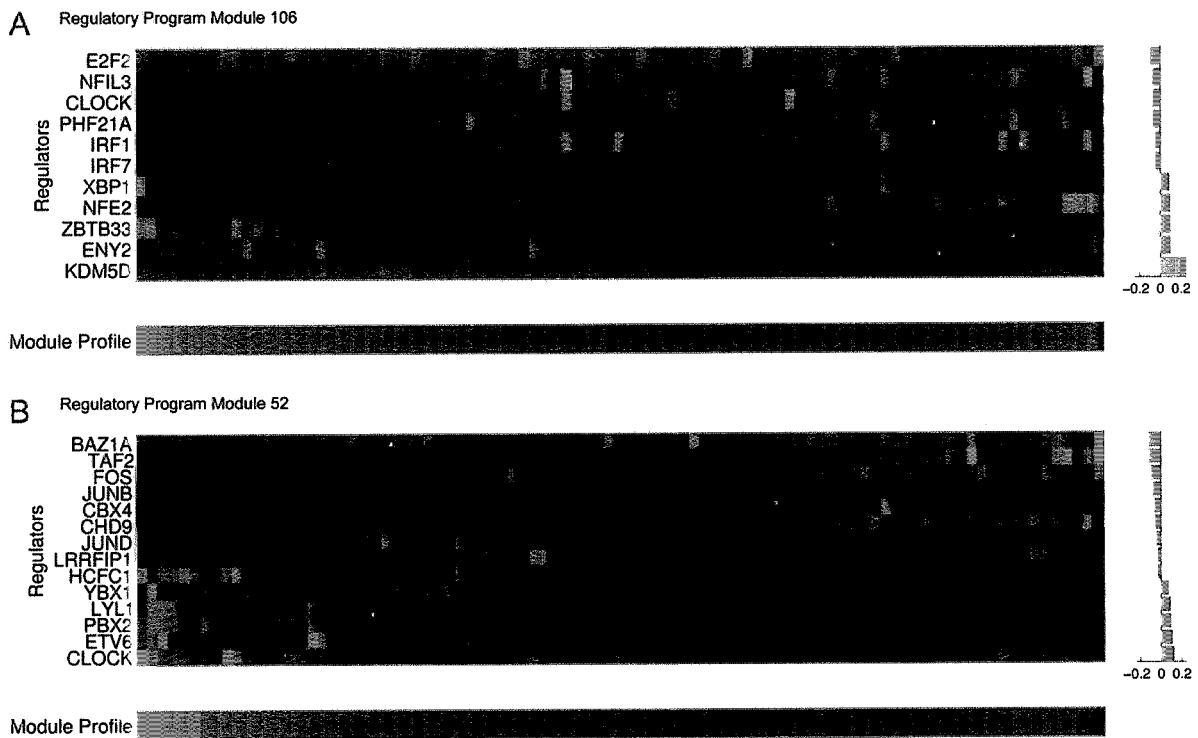

B   Regulatory Program Module 52

Fig. S1. Profile of gene regulators of modules 52 and 106. The initial gene expression data were reduced to gene modules by clustering analysis and assignment of a set of transcription factors (regulatory program) to each gene module. We used hierarchical agglomerative clustering to derive 109 modules. Using a set of candidate regulators composed of known signaling and transcription factors, for each gene module a set of regulatory genes (regulatory program) was assigned based on regression analysis of genes in the modules onto expression of transcription factors using the AIC (1). The regulatory program of module 52 contains FOS, JUNB, and JUD, among others, which is consistent with the suppressing effect of testosterone signaling on the AP-1 complex (FOS/JUN) (2). The regulatory program of sex-related gene module 106 contains transcription factors known to be differentially regulated in males versus females, such as CLOCK (3, 4), ENY2 (5), and IRF1 and IRF7 (6). Module profile, median expression of genes in the module.

1. Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Automat Contr* 19(6):716–723.
2. Kallio PJ, Poukka H, Moilanen A, Jänne OA, Palvimo JJ (1995) Androgen receptor-mediated transcriptional regulation in the absence of direct interaction with a specific DNA element. *Mol Endocrinol* 9(8):1017–1028.
3. Gómez-Abellán P, et al. (2012) Sexual dimorphism in clock genes expression in human adipose tissue. *Obes Surg* 22(1):105–112.
4. Lim AS, et al. (2013) Sex difference in daily rhythms of clock gene expression in the aged human cerebral cortex. *J Biol Rhythms* 28(2):117–129.
5. Xiao R, et al. (2012) In utero exposure to second-hand smoke aggravates adult responses to irritants: Adult second-hand smoke. *Am J Respir Cell Mol Biol* 47(6):843–851.
6. Haslinger C, et al. (2004) Microarray gene expression profiling of B-cell chronic lymphocytic leukemia subgroups defined by genomic aberrations and VH mutation status. *J Clin Oncol* 22(19):3937–3949.
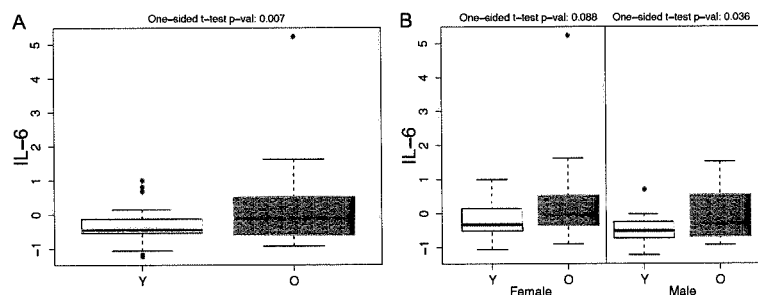
Fig. S2. Serum levels of IL-6 are increased in the elderly. To test for possibly explanatory variables of the elevated baseline levels of pSTAT3 proteins in blood monocytes in the elderly, we compared the serum IL-6 levels in all young (Y) versus older (O) individuals (A) or divided by sex (B). Significant differences are observed in all subjects (A), as well as in male individuals and to a lesser extent in females (B).
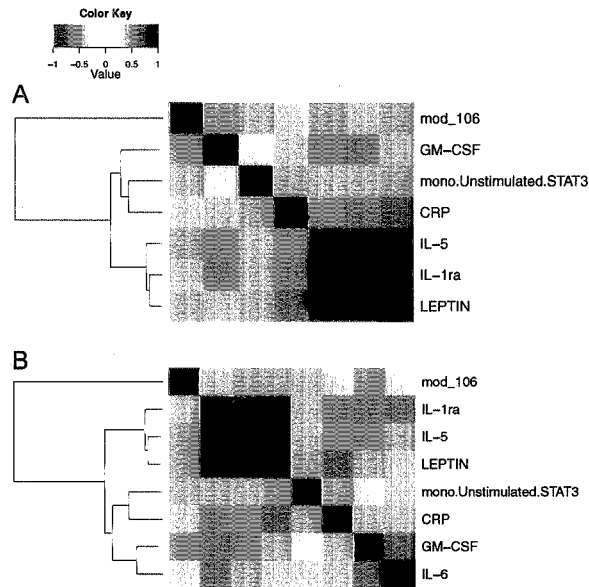
Fig. S3. Heat map of the correlation structure for sex-related immune features. A correlation matrix (Spearman method) was computed for all seven sex-related immune features in all individuals without (*A*) or with IL-6 (*B*), and hierarchical clustering (with Ward's method and Euclidian distance) was conducted. mono.Unstimulated.STAT3 clustered with CRP and to a lesser extent with GM-CSF (*A*, dendrogram), as well as with CRP and to a lesser extent with IL-6 and GM-CSF (*B*, dendrogram).
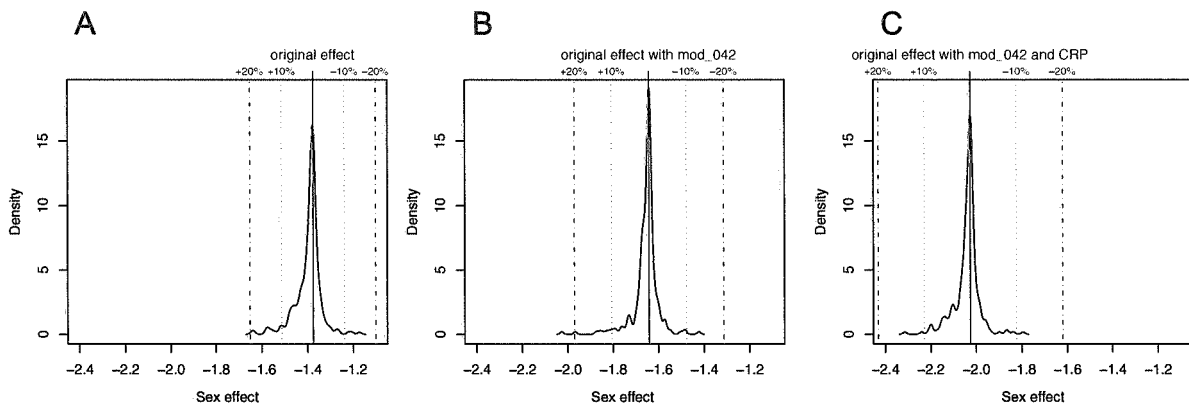
Fig. S4. Modifications in sex effect on vaccine response after adjusting for confounding factors. Forward stepwise logistic regression analysis was conducted to identify candidate confounders. (*A*) Regression coefficient of sex before adjustments. (*B*) Regression coefficient estimate for sex after adjusting for gene module 042. (*C*) Regression coefficient of sex after adjusting for gene module 042 and CRP levels. Negative values (*x* axis) indicate higher vaccine response in females.
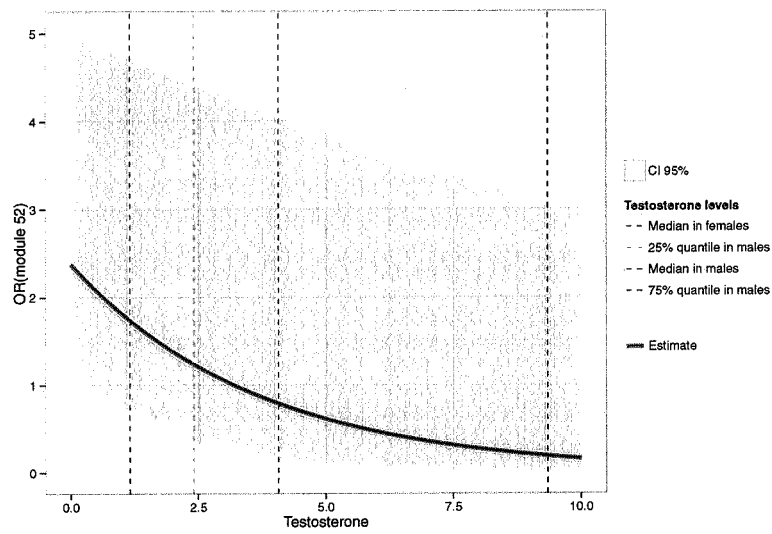
Fig. S5. Odds ratio for vaccine response based on expression of module 052 and testosterone levels. Logistic regression analysis was conducted on the antibody-neutralizing activity based on expression of genes in module 052 and the testosterone levels as a continuous measurement. The estimated odds ratio (OR) for the antibody-neutralizing response is shown (black continuous line). Red, light blue, blue, and dark blue dashed lines indicate the median testosterone levels in females and first, second, and third quartiles of testosterone levels for males. CI, confidence interval.

153

Table S1. Official gene symbol, Entrez ID, and module assignments for construction of gene modules 042, 052, and 106

Gene symbol	Entrez ID	Module assignment
RPS26P39	100128168	042
RPS26P38	100129552	042
RPS26P54	100131971	042
ZNF511	118472	042
SYT11	23208	042
GZMB	3002	042
RPS26P6	392256	042
RPS26P47	400156	042
ASCL2	430	042
RPS26P35	441377	042
RPS26P11	441502	042
ABI3	51225	042
CPSF3	51692	042
EXOSC10	5394	042
TRIT1	54802	042
PNPO	55163	042
RPS26	6231	042
RPS26P20	644166	042
RPS26P8	644191	042
RPS26P15	644928	042
RPS26P50	644934	042
RPS26P31	645979	042
RPS26P2	646753	042
RPS26P53	728823	042
RPS26P25	728937	042
CHRNA2	1135	052
FAM83F	113828	052
SPATA2L	124044	052
COX6C	1345	052
ZNF358	140467	052
CCDC140	151278	052
ADRA2C	152	052
AIM1	202	052
NAT9	26151	052
BSCL2	26580	052
GPR162	27239	052
C17orf60	284021	052
DHRS4L2	317749	052
HSPB1	3315	052
ANKRD33	341405	052
ARAF	369	052
FLJ41423	399886	052
RPS15P4	401019	052
LTA4H	4048	052
FAM116B	414918	052
MIF	4282	052
LOC440313	440313	052
LOC440993	440993	052
AURKAIP1	54998	052
USE1	55850	052
PDSS2	57107	052
PEX5	5830	052
RPS19	6223	052
BDKRB1	623	052
HSPBL2	653553	052
RPS19P3	728953	052
SPRYD3	84926	052
FIBCD1	84929	052
PIGQ	9091	052
NR1D1	9572	052
MTCP1NB	100272147	106
RPS4Y2	140032	106

154

Table S1. Cont.

Gene symbol	Entrez ID	Module assignment
CYorf15A	246126	106
NAAA	27163	106
NFU1	27247	106
GTF3A	2971	106
MTCP1	4515	106
PPA1	5464	106
PID1	55022	106
PLCXD1	55344	106
PRKY	5616	106
RPS4Y1	6192	106
CYorf15B	84663	106
ACCS	84680	106
DDX3Y	8653	106
EIF1AY	9086	106
KIAA0020	9933	106

155

Appendix F:

Gibbs sampler for Dirichlet process mixture of skew t-distributions models

- K is the number of different unique values taken by c (i.e. the number of clusters). This number of clusters K is not set and its value may change at each iteration.
- ℓ_c is the latent variable indicating which cluster the observation c belongs to. $\{\ell_{1:C}\}$ refers to a whole partition of the data.
- s_c is the skew parameter for the observation c.
- γ_c is the scale parameter (skew t only) for the observation c.

F.1 Skew Normal distributions mixture

Our Gibbs sampler proceeds with each of the following updates in turn:

1. update concentration parameter α given $\{\ell_{1:C}\}$ using the data augmentation technique from West [1992]:

   $\alpha \propto p(\alpha | \{z_{1:C}\}, G_0, \{\xi_k\}, \{\psi_k\}, \{\Sigma_k\}, \{\ell_{1:C}\}, \{w_k\}, \{s_{1:C}\}) \propto p(\alpha | \{\ell_{1:C}\})$

   $(\alpha, x | \{\ell_{1:C}\}) \sim p(\alpha) \alpha^{K-1} (\alpha + C) x^{\alpha} (1 - x)^{C-1}$

   $(x | \alpha, \{\ell_{1:C}\}) \sim \text{Beta}(\alpha + 1, C)$

   $(\alpha | x, \{\ell_{1:C}\}) \sim \pi_x \text{Gamma}(a + K, b - \log(x)) + (1 - \pi_x) \text{Gamma}(a + K - 1, b - \log(x))$

   with $p(\alpha) \propto \text{Gamma}(a, b)$ and $\frac{\pi_x}{1 - \pi_x} = \frac{a + k - 1}{C(b - \log(x))}$

2. update G given α, $\{\xi_k\}$, $\{\psi_k\}$, $\{\Sigma_k\}$ and G_0 via slice sampling:

   $\{w_k\}, \{\ell_{1:C}\} \propto p(\{w_k\}, \{\ell_{1:C}\} | \{z_{1:C}\}, \alpha, G_0, \{\xi_k\}, \{\psi_k\}, \{\Sigma_k\}, \{s_{1:C}\})$

   (a) sample the weights:

   $(w_1, \ldots, w_K, w_* | \{\ell_{1:C}\}) \sim \text{Dirichlet}(\text{card}(\{\ell_c = 1\}), \ldots, \text{card}(\{\ell_c = K\}), \alpha)$

   (b) for $c = 1, \ldots, C$: $u_c \sim \text{Unif}([0, w_{\ell_c}[)$

   (c) Set $j = K$. While $\sum_{k=1}^{j} w_k < (1 - \min(u_{1:C}))$:
   - set $j = j + 1$
   - sample $\pi_j \sim \text{Beta}(1, \alpha)$

— set $w_j = w_* \pi_j \prod_{k=K+1}^{j-1}(1 - \pi_k)$

— sample $(\xi_j, \psi_j, \Sigma_j | G_0) \sim G_0$

(d) for $c = 1, \ldots, C$ sample ℓ_c given $\{\xi_k\}, \{\psi_k\}, \{\Sigma_k\}, \{w_k\}$ from:

$$p(\ell_c = k) \propto \mathbb{1}_{\{w_k > u_c\}} f_{SN}(z_c, \xi_k, \psi_k, \Sigma_k)$$

3. for $c = 1, \ldots, C$ update s_c given ℓ_c , $\{\xi_k\}, \{\psi_k\}, \{\Sigma_k\}$:

$$p(s_c | z_c, \alpha, G_0, \{\xi_k\}, \{\psi_k\}, \{\Sigma_k\}, \{\ell_{1:C}\}, \{w_k\}) \propto p(s_c | z_c, \{\xi_k\}, \{\psi_k\}, \{\Sigma_k\}, \ell_c)$$

$$(s_c | z_c, \{\xi_k\}, \{\psi_k\}, \{\Sigma_k\}, \ell_c) \sim \mathcal{N}_{[0,+\infty[}(a_c, A_c)$$

with $A_c = \frac{1}{1 + \psi_{\ell_c}' \Sigma_{\ell_c}^{-1} \psi_{\ell_c}}$ and $a_c = A_c \psi_{\ell_c}' \Sigma_{\ell_c}^{-1}(z_c - \xi_{\ell_c})$

4. for $k = 1, \ldots, K$ update ξ_k, ψ_k and Σ_k given G_0, $\{\ell_{1:C}\}$ and $\{s_{1:C}\}$ from

$$p(\{\xi_k\}, \{\psi_k\}, \{\Sigma_k\} | \{z_{1:C}\}, \alpha, G_0, \{\ell_{1:C}\}, \{w_k\}, \{s_{1:C}\}):$$

(a) update G_k given $\{z_{1:C}\}$, G_0, $\{\ell_{1:C}\}$ and $\{s_{1:C}\}$:

— $G_0 = sNiW(b_0^\xi, b_0^\psi, B_0, \Lambda_0, \lambda_0)$ with $b_0 = (b_0^\xi{}' b_0^\psi{}')'$ and $B_0 = diag(D_0^\xi, D_0^\psi)$

— $G_k = sNiW(b_k^\xi, b_k^\psi, B_k, \Lambda_k, \lambda_k)$ with $b_k = (b_k^\xi{}' b_k^\psi{}')'$

— let X_k be a matrix of dimension $card(\{c | \ell_c = k\}) \times 2$: $X_k = (1 \; s_{c | \ell_c = k})$

— let $B_k = (X_k' X_k + diag(D_0)^{-1})^{-1}$

— $b_k = \left(z_{c | \ell_c = k} X_k + \left(\frac{1}{D_0^\xi} b_0^\xi \; \frac{1}{D_0^\psi} b_0^\psi\right)\right) B_k$

— $\lambda_k = \lambda_0 + card(\{c | \ell_c = k\})$

— $\Lambda_k = \Lambda_0 + \sum_{c | \ell_c = k} \varepsilon_c \varepsilon_c' + \frac{1}{D_0^\xi}(b_k^\xi - b_0^\xi)(b_k^\xi - b_0^\xi)' + \frac{1}{D_0^\psi}(b_k^\psi - b_0^\psi)(b_k^\psi - b_0^\psi)$

with $\varepsilon_c = z_c - b_k^\xi - s_c b_k^\psi$

(b) sample $(\xi_k, \psi_k, \Sigma_k | G_k) \sim G_k$

— $((\xi_k, \psi_k) | \Sigma_k, \{\ell_{1:C}\}, \{s_{1:C}\}, G_k) \sim \mathcal{N}_{2d}\left((b_k^\xi, b_k^\psi), B_k \otimes \Sigma_k\right)$

— $(\Sigma_k | \{\ell_{1:C}\}, \{s_{1:C}\}, G_k) \sim \mathcal{W}^{-1}(\lambda_k, \Lambda_k)$

F.2 Skew *t*-distributions mixture

Our Gibbs sampler for non parametric skew *t*-distributions mixture proceeds with each of the following updates in turn:

1. update concentration parameter α given $\{\ell_{1:C}\}$ using the data augmentation technique from West [1992]:

$\alpha \propto p(\alpha | \{z_{1:C}\}, G_0, \{\xi_k\}, \{\psi_k\}, \{\Sigma_k\}, \{\nu_k\}, \{\ell_{1:C}\}, \{w_k\}, \{s_{1:C}\}, \{\gamma_{1:C}\}) \propto p(\alpha | \{\ell_{1:C}\})$

$(\alpha, x | \{\ell_{1:C}\}) \sim p(\alpha) \alpha^{K-1}(\alpha + C) x^\alpha (1 - x)^{C-1}$

$(x | \alpha, \{\ell_{1:C}\}) \sim Beta(\alpha + 1, C)$

$(\alpha | x, \{\ell_{1:C}\}) \sim \pi_x Gamma(a + K, b - \log(x)) + (1 - \pi_x)Gamma(a + K - 1, b - \log(x))$

with $p(\alpha) \propto Gamma(a, b)$ and $\frac{\pi_x}{1 - \pi_x} = \frac{a + k - 1}{C(b - \log(x))}$

2. update G given α, $\{\xi_k\}$, $\{\psi_k\}$, $\{\Sigma_k\}$, $\{\nu_k\}$ and G_0 via slice sampling:

$$\{w_k\}, \{\ell_{1:C}\} \propto p(\{w_k\}, \{\ell_{1:C}\} | \{z_{1:C}\}, \alpha, G_0, \{\xi_k\}, \{\psi_k\}, \{\Sigma_k\}, \{\nu_k\}, \{s_{1:C}\}, \{\gamma_{1:C}\})$$

   (a) sample the weights:

$$(w_1, \ldots, w_K, w_* | \{\ell_{1:C}\}) \sim \text{Dirichlet}\,(\text{card}(\{\ell_{1:C}\} = 1), \ldots, \text{card}(\{\ell_{1:C}\} = K), \alpha)$$

   (b) for $c = 1, \ldots, C$: $u_c \sim \text{Unif}([0, w_{\ell_c}])$

   (c) Set $j = K$. While $\sum_{k=1}^{j} w_k < (1 - min(u_{1:C}))$:

      — set $j = j + 1$

      — sample $\pi_j \sim \text{Beta}(1, \alpha)$

      — set $w_j = w_* \pi_j \prod_{k=K+1}^{j-1}(1 - \pi_k)$

      — sample $(\xi_j, \psi_j, \Sigma_j | G_0) \sim \text{structured-Normal-invWishart}(G_0)$

      — sample $\nu_j \sim p(\nu_j)$

   (d) $K = j$

   (e) for $c = 1, \ldots, C$ sample ℓ_c given $\{\xi_k\}$, $\{\psi_k\}$, $\{\Sigma_k\}$, $\{w_k\}$ from:

$$p(\ell_c = k) \propto \mathbb{1}_{\{w_k > u_c\}} f_{SN}(z_c, \xi_k, \psi_k, \Sigma_k)$$

3. for $c = 1, \ldots, C$ update s_c given ℓ_c, $\{\xi_k\}$, $\{\psi_k\}$, $\{\Sigma_k\}$:

$$(s_c | z_c, \{\xi_k\}, \{\psi_k\}, \{\Sigma_k\}, \ell_c) \sim \mathcal{N}_{[0,+\infty[}(a_c, A_c)$$

with $A_c = \frac{1}{1 + \psi'_{\ell_c} \Sigma_{\ell_c}^{-1} \psi_{\ell_c}}$ and $a_c = A_c \psi'_{\ell_c} \Sigma_{\ell_c}^{-1}(z_c - \xi_{\ell_c})$

4. for $k = 1, \ldots, K$ update ξ_k, ψ_k and Σ_k given G_0, $\{\ell_{1:C}\}$ and $\{s_{1:C}\}$ from:

$$p(\{\xi_k\}, \{\psi_k\}, \{\Sigma_k\} | \{z_{1:C}\}, \alpha, G_0, \{\nu_k\}, \{\ell_{1:C}\}, \{w_k\}, \{s_{1:C}\}, \{\gamma_{1:C}\})$$
$$\propto p(\{\xi_k\}, \{\psi_k\}, \{\Sigma_k\} | \{\ell_{1:C}\}, \{s_{1:C}\}, G_0):$$

   (a) update the hyper parameters of the cluster distribution given $\{z_{1:C}\}$, G_0, $\{\ell_{1:C}\}$ and $\{s_{1:C}\}$:

      — $G_0 = sNiW(b_0^\xi, b_0^\psi, B_0, \Lambda_0, \lambda_0)$ with $b_0 = vec(b_0^\xi, b_0^\psi)$ and $B_0 = diag(D_0^\xi, D_0^\psi)$

      — $G_k = sNiW(b_k^\xi, b_k^\psi, B_k, \Lambda_k, \lambda_k)$ with $b_k = vec(b_k^\xi, b_k^\psi)$

      — let X_k be a matrix of dimension $\text{card}(\{c | \ell_c = k\}) \times 2$: $X_k = (1 \ \ s_{c|\ell_c=k})$

      — let $B_k = (X'_k X_k + (B_0)^{-1})^{-1}$

      — $b_k = \left(z_{c|\ell_c=k} X_k + \left(\frac{1}{D_0^\xi} b_0^\xi \ \ \frac{1}{D_0^\psi} b_0^\psi\right)\right) B_k$

      — $\lambda_k = \lambda_0 + \text{card}(\{\ell_c = k\})$

      — $\Lambda_k = \Lambda_0 + \sum_{c|\ell_c=k} \varepsilon_c \varepsilon'_c + \frac{1}{D_0^\xi}(b_k^\xi - b_0^\xi)(b_k^\xi - b_0^\xi)' + \frac{1}{D_0^\psi}(b_k^\psi - b_0^\psi)(b_k^\psi - b_0^\psi)$

      with $\varepsilon_c = z_c - b_k^\xi - s_c b_k^\psi$

   (b) sample $(\xi_k, \psi_k, \Sigma_k | b_k, B_k, \Lambda_k, \lambda_k)$ from a $sNiW(b_k, B_k, \Lambda_k, \lambda_k)$

      — $((\xi_k, \psi_k) | \Sigma_k, \{\ell_{1:C}\}, \{s_{1:C}\}, G_k) \sim \mathcal{N}_{2d}\left((b_k^\xi, b_k^\psi), B_k \otimes \Sigma_k\right)$

      — $(\Sigma_k | \{\ell_{1:C}\}, \{s_{1:C}\}, G_k) \sim \mathcal{W}^{-1}(\lambda_k, \Lambda_k)$

5. update the degrees of freedom $\{\nu_k\}$ and the scale factors $\{\gamma_{1:C}\}$ from the random effects representation given $\{\boldsymbol{\xi}_k\}$, $\{\boldsymbol{\psi}_k\}$, $\{\boldsymbol{\Sigma}_k\}$, $\{s_{1:C}\}$ and $\{\ell_{1:C}\}$, sampling from:

$p(\nu_k, \{\gamma_{1:C}\} | \{\boldsymbol{\xi}_k\}, \{\boldsymbol{\psi}_k\}, \{\boldsymbol{\Sigma}_k\}, \{s_{1:C}\}, \{\ell_{1:C}\})$

(a) for $k = 1, \ldots, K$ update ν_k, given $\boldsymbol{\xi}_k$, $\boldsymbol{\psi}_k$, $\boldsymbol{\Sigma}_k$, $\{s_{1:C}\}$ and $\{\ell_{1:C}\}$, integrating out the $\{\gamma_{1:C}\}$, sampling from:

$p(\nu_k | \{\boldsymbol{\xi}_k\}, \{\boldsymbol{\psi}_k\}, \{\boldsymbol{\Sigma}_k\}, \{\ell_{1:C}\}, \{w_k\}, \{s_{1:C}\}, \alpha, \{\gamma_{1:C}\})$

$\propto p(\nu_k | \{\boldsymbol{\xi}_k\}, \{\boldsymbol{\psi}_k\}, \{\boldsymbol{\Sigma}_k\}, \{\ell_{1:C}\}, \{s_{1:C}\}, \{\gamma_{1:C}\})$

$\propto p(\nu_k | \{\boldsymbol{\xi}_k\}, \{\boldsymbol{\psi}_k\}, \{\boldsymbol{\Sigma}_k\}, \{\ell_{1:C}\}, \{s_{1:C}\})$ (reducing conditioning on the $\{\gamma_{1:C}\}$)

A Metropolis-Hastings step is required to sample from the above distribution. We use a uniform log random-walk proposal as proposed in Frühwirth-Schnatter and Pyne [2010]:

$$\log(\nu_k^{new} - 1) \sim \text{Unif}([\log(\nu_k - 1) - c_{\nu_k}, \log(\nu_k - 1) + c_{\nu_k}]$$

where c_{ν_k} is a fixed parameter of the algorithm (that can be tuned to improve the acceptance rate of this MH step). Acceptance probability for ν_k^{new} is as follow:

$$\min\left(1, \frac{p(y|\{\boldsymbol{\xi}_k\}, \{\boldsymbol{\psi}_k\}, \{\boldsymbol{\Sigma}_k\}, \nu_{-k}, \nu_k^{new}, \{\ell_{1:C}\}) p(\nu_k^{new})(\nu_k^{new} - 1)}{p(y|\{\boldsymbol{\xi}_k\}, \{\boldsymbol{\psi}_k\}, \{\boldsymbol{\Sigma}_k\}, \{\nu_k\}, \{\ell_{1:C}\}) p(\nu_k)(\nu_k^{-}1)}\right)$$

(b) for $c = 1, \ldots, C$ update γ_c given $\{\boldsymbol{\xi}_k\}$, $\{\boldsymbol{\psi}_k\}$, $\{\boldsymbol{\Sigma}_k\}$, $\{\nu_k\}$, s_c and ℓ_c sampling from:

$$p(\gamma_c | \{\boldsymbol{\xi}_k\}, \{\boldsymbol{\psi}_k\}, \{\boldsymbol{\Sigma}_k\}, \{\nu_k\}, s_c, \ell_c) \sim \text{Gamma}\left(\frac{\nu_{\ell_c} + d + 1}{2}, \frac{\nu_{\ell_c} + z_c^2 + tr(\eta_c \eta_c' \Sigma_{\ell_c}^{-1})}{2}\right)$$

with $\eta_c = \boldsymbol{z}_c - \boldsymbol{\xi}_{\ell_c} - s_c \boldsymbol{\psi}_{\ell_c}$

F.3 Skew t-distributions mixture with informative mixture of priors

Now we consider the case where the prior G_0 is actually a mixture of different priors given by :

$$G_0 = \sum_{j=1}^{J} sNiW(\boldsymbol{b}_{0j}^{\xi}, \boldsymbol{b}_{0j}^{\psi}, \boldsymbol{B}_{0j}, \Lambda_{0j}, \lambda_{0j})$$

Let $\boldsymbol{U}_k = (\boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k)$. The base measure over the cluster locations is then

$$p(\boldsymbol{U}_k, \nu_k) = p(\boldsymbol{U}_k) p(\nu_k)$$

$$= \left(\sum_{j=1}^{J} \omega_j f_j(\boldsymbol{U}_k)\right) p(\nu_k)$$

where f_j is a structured Normal inverse Wishart distribution of parameters $\phi_j = \{b_{0j}^\xi,\ b_{0j}^\psi,\ B_{0j}, \Lambda_{0j}, \lambda_{0j}\}$:

$$f_j(\cdot) = sNiW(\cdot; b_{0j}^\xi, b_{0j}^\psi, B_{0j}, \Lambda_{0j}, \lambda_{0j})$$

Let y_{A_k} where $A_k = \{c|\ell_c = k\}$ is the data in a given cluster k. Conditional on ν_k, we have

$$p(U_k|y_{A_k}, \nu_k) = \frac{p(y_{A_k}|U_k, \nu_k)p(U_k|\nu_k)}{p(y_k|\nu_k)}$$
$$= \frac{p(y_{A_k}|U_k, \nu_k)p(U_k)}{\int p(y_{A_k}|U, \nu_k)p(U)dU}$$

Now

$$p(y_{A_k}|U_k, \nu_k)p(U_k) = \sum_{j=1}^{J} \omega_j p(y_{A_k}|U_k, \nu_k)f_j(U_k)$$

Let's write

$$p_j(U_k|y_{A_k}) = \frac{p(y_{A_k}|U_k, \nu_k)f_j(U_k)}{\int p(y_{A_k}|U, \nu_k)f_j(U)dU}$$

and $\quad p_j(y_{A_k}) = \int p(y_{A_k}|U, \nu_k)f_j(U)dU = \dfrac{p(y_{A_k}|U_k, \nu_k)f_j(U_k)}{p_j(U_k|y_{A_k})}$

As we are in a conjugate setting, $p_j(U_k|y_{A_k})$ is a sNiW whose probability density function can be evaluated analytically, and $p_j(y_{A_k})$ can be evaluated analytically. Thus we have

$$p(U_k|y_{A_k}, \nu_k) = \sum_{j=1}^{J} \frac{\omega_j p_j(y_{A_k})}{\sum_{j'=1}^{J} \omega_{j'} p_{j'}(y_{A_k})} p_j(U_k|y_{A_k})$$

So the update for each cluster k in step 4 of the sampler becomes:

(a) For each $j = 1, \ldots, J$, compute the sufficient statistics b_{jk}, B_{jk}, Λ_{kj}, λ_{jk}, and the associated $p_j(y_{A_k})$

(b) Sample an index $m \in \{1, \ldots, J\}$ from the discrete distribution

$$\left(\frac{\omega_1 p_1(y_{A_k})}{\sum_{j'=1}^{J} \omega_{j'} p_{j'}(y_{A_k})}, \ldots, \frac{\omega_J p_J(y_{A_k})}{\sum_{j'=1}^{J} \omega_{j'} p_{j'}(y_{A_k})} \right)$$

(c) Sample $U_k|m \sim p_m(U_k|y_{A_k})$

The update for ν_k in step 5 remains the same.

F.4   MH within collapsed Gibbs

As an MH is used in the skew t sampler to sample $\{\nu_k\}$, it is very important to never integrate out those $\{\nu_k\}$ in the previous steps of the Partially Collapsed Gibbs sampler [van Dyk and Jiao, 2014]. Otherwise, there is no guaranty that the stationary distribution of the Markov chain remains unchanged (correlation structure of the $\{\nu_k\}$ with the other parameters is likely not to be estimated properly). Besides, the reduced conditioning on the $\{\gamma_{1:C}\}$ does not change the stationary distribution as those marginalized out $\{\gamma_{1:C}\}$ are sampled right after the MH step from their full conditional distribution [van Dyk and Jiao, 2014].

Appendix G:

Parameter estimation for Normal inverse-Wishart and structured Normal inverse-Wishart distributions

G.1 Maximum Likelihood Estimation

G.1.1 Maximum Likelihood estimators for Normal inverse-Wishart

Let observations (μ_i, Σ_i) follow a Normal inverse-Wishart distribution for $i = 1 \ldots n$:

$$(\mu_i, \Sigma_i) \sim NiW(\mu_0, \kappa_0, \Lambda_0, \lambda_0)$$

The likelihood is:

$$p\left(\{\mu_{1:n}\}, \{\Sigma_{1:n}\} | \mu_0, \kappa_0, \Lambda_0, \lambda_0\right) = \prod_{i=1}^{n} \left\{ (2\pi)^{-\frac{d}{2}} |\Sigma_i|^{-\frac{\lambda_0+d+1}{2}} \frac{2^{-\frac{\lambda_0 d}{2}} |\Lambda_0|^{\frac{\lambda_0}{2}}}{\Gamma_d(\frac{\lambda_0}{2})} \left| \frac{1}{\kappa_0} \Sigma_i \right|^{-\frac{1}{2}} \right.$$

$$\left. \exp\left[-\frac{1}{2} tr\left(\Lambda_0 \Sigma_i^{-1}\right) - \frac{\kappa_0}{2} (\mu_i - \mu_0)' \Sigma_i^{-1} (\mu_i - \mu_0) \right] \right\}$$

The loglikelihood is then:

$$\log\left(p\left(\{\mu_{1:n}\}, \{\Sigma_{1:n}\} | \mu_0, \kappa_0, \Lambda_0, \lambda_0\right)\right) = -\frac{d}{2}\log(2\pi) - \frac{\lambda_0+d+2}{2} \sum_{i=1}^{n} \log\left(|\Sigma_i|\right) - \frac{n\lambda_0 d}{2}\log(2)$$

$$+ \frac{n\lambda_0}{2}\log\left(|\Lambda_0|\right) - n\log\left(\Gamma_d(\frac{\lambda_0}{2})\right) + \frac{nd}{2}\log\left(\kappa_0\right)$$

$$- \frac{1}{2} tr\left(\Lambda_0 \sum_{i=1}^{n} \Sigma_i^{-1}\right) - \frac{\kappa_0}{2} \sum_{i=1}^{n} (\mu_i - \mu_0)' \Sigma_i^{-1} (\mu_i - \mu_0)$$

Taking the partial derivatives in respect of the four parameters $\mu_0, \kappa_0, \Lambda_0, \lambda_0$ and setting

163

each of them to zero gives the following system:

$$\begin{cases} \mu_0 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\mu_i \\[2ex] \dfrac{1}{\kappa_0} = \dfrac{1}{nd}\displaystyle\sum_{i=1}^{n}(\mu_i - \mu_0)'\Sigma_i^{-1}(\mu_i - \mu_0) \\[2ex] \Lambda_0 = n\lambda_0\left(\displaystyle\sum_{i=1}^{n}\Sigma_i^{-1}\right)^{-1} \\[2ex] 0 = -\dfrac{1}{2}\displaystyle\sum_{i=1}^{n}\log\left(|\Sigma_i|\right) - \dfrac{nd}{2}\log(2) + \dfrac{n}{2}\log\left(|\Lambda_0|\right) - \dfrac{n}{2}F_d\left(\dfrac{\lambda_0}{2}\right) \end{cases}$$

where $F_d(x) = \dfrac{\mathrm{d}}{\mathrm{d}x}\log(\Gamma_d(x))$ is the d-dimensional digamma function (the derivative of the logarithm of the d-dimensional Gamma function).

NB: The above solution are obtained using the two following identities: $\dfrac{\mathrm{d}}{\mathrm{d}X}\log(|X|) = X^{-1}$ and $\dfrac{\mathrm{d}}{\mathrm{d}X}tr(XA) = A'$ if X is definite-positive

Hence the MLE solutions verify:

$$\begin{cases} \widehat{\mu_0} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\mu_i \\[2ex] \widehat{\kappa_0} = nd\left(\displaystyle\sum_{i=1}^{n}(\mu_i - \widehat{\mu_0})'\Sigma_i^{-1}(\mu_i - \widehat{\mu_0})\right)^{-1} \\[2ex] F_d\left(\dfrac{\widehat{\lambda_0}}{2}\right) = -\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\log\left(|\Sigma_i|\right) + d\log\left(\dfrac{n\widehat{\lambda_0}}{2}\right) - \log\left(\left|\displaystyle\sum_{i=1}^{n}\Sigma_i^{-1}\right|\right) \\[2ex] \widehat{\Lambda_0} = n\widehat{\lambda_0}\left(\displaystyle\sum_{i=1}^{n}\Sigma_i^{-1}\right)^{-1} \end{cases}$$

under the constraint $\widehat{\lambda_0} > d + 1$ (in which case there should a unique solution $\widehat{\lambda_0}$).

G.1.2 Maximum Likelihood estimators for structured Normal inverse-Wishart

Let observations $(\xi_i, \psi_i, \Sigma_i)$ follow a structured Normal inverse-Wishart distribution $(sNiW)$ for $i = 1\ldots n$:

$$(\xi_i, \psi_i, \Sigma_i) \sim sNiW\left(\xi_0, \psi_0, B_0, \Lambda_0, \lambda_0\right)$$

The likelihood is:

$$p\left(\{\boldsymbol{\xi}_{1:n}\}, \{\boldsymbol{\psi}_{1:n}\}, \{\boldsymbol{\Sigma}_{1:n}\}|\boldsymbol{\mu}_0, \boldsymbol{B}_0, \boldsymbol{\Lambda}_0, \lambda_0\right) = \prod_{i=1}^{n} \left\{ (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}_i|^{-\frac{\lambda_0+d+1}{2}} \frac{2^{-\frac{\lambda_0 d}{2}} |\boldsymbol{\Lambda}_0|^{\frac{\lambda_0}{2}}}{\Gamma_d(\frac{\lambda_0}{2})} \left|\boldsymbol{B}_0^{-1} \otimes \boldsymbol{\Sigma}_i\right|^{-\frac{1}{2}} \right.$$

$$\left. \exp\left[-\frac{1}{2} tr\left(\boldsymbol{\Lambda}_0 \boldsymbol{\Sigma}_i^{-1}\right) \right. \right.$$

$$\left. \left. -\frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)'\left(\boldsymbol{B}_0 \otimes \boldsymbol{\Sigma}_i^{-1}\right)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)\right]\right\}$$

where $\boldsymbol{\mu}_i = (\boldsymbol{\xi}_i' \; \boldsymbol{\psi}_i')'$ and $\boldsymbol{\mu}_0 = (\boldsymbol{\xi}_0' \; \boldsymbol{\psi}_0')'$

The loglikelihood is then:

$$\log\left(p\left(\{\boldsymbol{\mu}_{1:n}\}, \{\boldsymbol{\Sigma}_{1:n}\}|\boldsymbol{\mu}_0, \boldsymbol{B}_0, \boldsymbol{\Lambda}_0, \lambda_0\right)\right) = -\frac{nd}{2}\log(2\pi) - \frac{\lambda_0+d+1}{2}\sum_{i=1}^{n}\log\left(|\boldsymbol{\Sigma}_i|\right) - \frac{n\lambda_0 d}{2}\log(2)$$

$$+ \frac{n\lambda_0}{2}\log\left(|\boldsymbol{\Lambda}_0|\right) - n\log\left(\Gamma_d(\frac{\lambda_0}{2})\right)$$

$$- \frac{1}{2}\sum_{i=1}^{n}\log\left(\left|\boldsymbol{B_0}^{-1} \otimes \boldsymbol{\Sigma}_i\right|\right) - \frac{1}{2}tr\left(\boldsymbol{\Lambda}_0\sum_{i=1}^{n}\boldsymbol{\Sigma}_i^{-1}\right)$$

$$- \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)'\left(\boldsymbol{B_0} \otimes \boldsymbol{\Sigma}_i^{-1}\right)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)$$

Taking the partial derivatives in respect of the four parameters $\mu_0, B_0, \Lambda_0, \lambda_0$ and setting each of them to zero gives the following system:

$$\begin{cases} \boldsymbol{\mu}_0 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\boldsymbol{\mu}_i \\[2mm] 0 = \displaystyle\sum_{i=1}^{n}\left(\dfrac{\mathrm{d}}{\mathrm{d}\boldsymbol{B_0}}\left(\log\left(\left|\boldsymbol{B_0}^{-1} \otimes \boldsymbol{\Sigma}_i\right|\right)\right) + \dfrac{\mathrm{d}}{\mathrm{d}\boldsymbol{B_0}}\left((\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)'\left(\boldsymbol{B_0} \otimes \boldsymbol{\Sigma}_i^{-1}\right)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)\right)\right) \\[2mm] \boldsymbol{\Lambda}_0 = n\lambda_0\left(\displaystyle\sum_{i=1}^{n}\boldsymbol{\Sigma}_i^{-1}\right)^{-1} \\[2mm] 0 = -\dfrac{1}{2}\displaystyle\sum_{i=1}^{n}\log\left(|\boldsymbol{\Sigma}_i|\right) - \dfrac{nd}{2}\log(2) + \dfrac{n}{2}\log\left(|\boldsymbol{\Lambda}_0|\right) - \dfrac{n}{2}F_d\left(\dfrac{\lambda_0}{2}\right) \end{cases}$$

where $F_d(x) = \dfrac{\mathrm{d}}{\mathrm{d}x}\log(\Gamma_d(x))$ is the digamma function (the derivative of the logarithm of the Gamma function).

$$\sum_{i=1}^{n} \left(\frac{\mathrm{d}}{\mathrm{d}B_0} \left(\log \left(|B_0^{-1} \otimes \Sigma_i| \right) \right) + \frac{\mathrm{d}}{\mathrm{d}B_0} \left((\mu_i - \mu_0)' \left(B_0 \otimes \Sigma_i^{-1} \right) (\mu_i - \mu_0) \right) \right)$$

$$= \sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}B_0} \left(\log \left(|B_0|^{-d} |\Sigma_i|^2 \right) \right) + \sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}B_0} \left((\mu_i - \mu_0)' \left(B_0 \otimes \Sigma_i^{-1} \right) (\mu_i - \mu_0) \right)$$

$$= -nd \frac{\mathrm{d}}{\mathrm{d}B_0} \left(\log \left(|B_0| \right) \right) + \sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}B_0} \left((\mu_i - \mu_0)' \left(B_0 \otimes \Sigma_i^{-1} \right) (\mu_i - \mu_0) \right)$$

$$= -nd B_0^{-1} + \sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}B_0} \left((\mu_i - \mu_0)' \left(B_0 \otimes \Sigma_i^{-1} \right) (\mu_i - \mu_0) \right)$$

$$= -nd B_0^{-1} + \sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}B_0} \left(tr \left((\mu_i - \mu_0)' \left(B_0 \otimes \Sigma_i^{-1} \right) (\mu_i - \mu_0) \right) \right)$$

$$= -nd B_0^{-1} + \sum_{i=1}^{n} \left(\begin{array}{cc} \xi_i - \xi_0 & \psi_i - \psi_0 \end{array} \right)' \left(\Sigma_i^{-1} \right) \left(\begin{array}{cc} \xi_i - \xi_0 & \psi_i - \psi_0 \end{array} \right)$$

$$= -nd B_0^{-1} + \sum_{i=1}^{n} \left(\begin{array}{c} \xi_i' - \xi_0' \\ \psi_i' - \psi_0' \end{array} \right) \left(\Sigma_i^{-1} \right) \left(\begin{array}{cc} \xi_i - \xi_0 & \psi_i - \psi_0 \end{array} \right)$$

So if the above expression is zero, we get:

$$\widehat{B_0} = nd \left(\sum_{i=1}^{n} \left(\begin{array}{c} \xi_i' - \xi_0' \\ \psi_i' - \psi_0' \end{array} \right) \left(\Sigma_i^{-1} \right) \left(\begin{array}{cc} \xi_i - \xi_0 & \psi_i - \psi_0 \end{array} \right) \right)^{-1}$$

So MLE solution for $sNiW$ are:

$$\begin{cases} \widehat{\xi_0} = \frac{1}{n} \sum_{i=1}^{n} \xi_i \\[2mm] \widehat{\psi_0} = \frac{1}{n} \sum_{i=1}^{n} \psi_i \\[2mm] \widehat{B_0} = nd \left(\sum_{i=1}^{n} \left(\begin{array}{c} \xi_i' - \xi_0' \\ \psi_i' - \psi_0' \end{array} \right) \left(\Sigma_i^{-1} \right) \left(\begin{array}{cc} \xi_i - \xi_0 & \psi_i - \psi_0 \end{array} \right) \right)^{-1} \\[2mm] F_d \left(\frac{\widehat{\lambda_0}}{2} \right) = -\frac{1}{n} \sum_{i=1}^{n} \log \left(|\Sigma_i| \right) + d \log \left(\frac{n \widehat{\lambda_0}}{2} \right) - \log \left(\left| \sum_{i=1}^{n} \Sigma_i^{-1} \right| \right) \\[2mm] \widehat{\Lambda_0} = n \widehat{\lambda_0} \left(\sum_{i=1}^{n} \Sigma_i^{-1} \right)^{-1} \end{cases}$$

G.2 Expectation-Maximization algorithms (MLE & MAP)

G.2.1 MLE estimation via an E-M algorithm

The latent variables used in the EM [Dempster et al., 1977] algorithm for estimating a finite mixture model over the MCMC draws for the parameters $\boldsymbol{\xi}_i$, $\boldsymbol{\psi}_i$ and Σ_i are the allocation variables ℓ_i, with $i = 1..N$ the number of (MCMC) observations. An (MCMC) observation is then $\boldsymbol{x}_i = (\boldsymbol{\xi}_i, \boldsymbol{\psi}_i, \Sigma_i)$. Let K be the number of components in the mixture model:

$$p(\boldsymbol{x}_i | K, \{\boldsymbol{\theta}_{1:K}\}) = \sum_{k=1}^{K} \pi_k f_{\boldsymbol{\theta}_{\ell_i}}(\boldsymbol{x}_i | \ell_i, \{\boldsymbol{\theta}_{1:K}\}) \qquad \text{for } i = 1 \ldots N$$

where $f_{\boldsymbol{\theta}_k}$ is the parametric density function of a cluster: a $sNiW$ density function with parameters $\boldsymbol{\theta}_k = (\boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{B}_k, \Lambda_k, \lambda_k)$.

At iteration t, the EM algorithm maximizes $Q\left(\{\boldsymbol{\theta}_{1:K}\} \middle| \{\boldsymbol{\theta}_{1:K}^{(t-1)}\}\right)$ for $\{\boldsymbol{\theta}_{1:K}\}$ with:

$$
\begin{aligned}
Q\left(\{\boldsymbol{\theta}_{1:K}\} \middle| \{\boldsymbol{\theta}_{1:K}^{(t-1)}\}\right) &= \mathbb{E}\left[\log\left(p(\boldsymbol{x}_{\{1:n\}}, \ell_{\{1:N\}} | K, \{\boldsymbol{\theta}_{1:K}\})\right) \middle| \{\boldsymbol{\theta}_{1:K}^{(t-1)}\}\right] \\
&= \sum_{\ell_{\{1:N\}}} \log\left(p(\boldsymbol{x}_{\{1:n\}}, \ell_{\{1:N\}} | K, \{\boldsymbol{\theta}_{1:K}^{(t-1)}\})\right) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{n} r_{ik}^{(t-1)} \log(\pi_k) + \sum_{k=1}^{K}\sum_{i=1}^{n} r_{ik}^{(t-1)} \log\left(p(\boldsymbol{x}_i | K, \{\boldsymbol{\theta}_{1:K}\})\right) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{n} \Bigg[r_{ik}^{(t-1)} \log(\pi_k) - \frac{\lambda_k + d + 1}{2} r_{ik}^{(t-1)} \log(|\Sigma_i|) \\
&\quad - \frac{\lambda_k d r_{ik}^{(t-1)}}{2} \log(2) - r_{ik}^{(t-1)} \log(\Gamma_d(\frac{\lambda_k}{2})) + \frac{r_{ik}^{(t-1)}\lambda_k}{2} \log(|\Lambda_k|) \\
&\quad - \frac{r_{ik}^{(t-1)}}{2} \log(|\boldsymbol{B}_k^{-1} \otimes \Sigma_i|) - \frac{r_{ik}^{(t-1)}}{2} tr(\Lambda_k \Sigma_i^{-1}) \\
&\quad - \frac{r_{ik}^{(t-1)}}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_k)'(\boldsymbol{B}_k \otimes \Sigma_i^{-1})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_k) \Bigg]
\end{aligned}
$$

with $r_{ik}^{(t)} = p(\ell_i = k | \boldsymbol{x}_i, \{\boldsymbol{\theta}_{1:K}^{(t)}\}) = \dfrac{\pi_k f_{\boldsymbol{\theta}_k^{(t)}}(\boldsymbol{x}_i)}{\sum_{j=1}^{K} \pi_j f_{\boldsymbol{\theta}_j^{(t)}}(\boldsymbol{x}_i)}$

1. **Initialization**

   $\boldsymbol{\theta}_k^{(0)}$ is initialized randomly (π_k are initialized at $1/K$)

2. **E step at iteration t**

   Compute the membership weights $r_{ik}^{(t-1)}$ for each observation $i = 1 \ldots N$ for each cluster $k = 1 \ldots K$:

$$r_{ik}^{(t-1)} = p\left(\ell_i = k \,\middle|\, x_i, \{\theta_{1:K}^{(t-1)}\}\right) = \frac{\pi_k f_{\theta_k^{(t-1)}}(x_i)}{\sum_{j=1}^{K} \pi_j f_{\theta_j^{(t-1)}}(x_i)}$$

3. **M step at iteration** t

Update the parameters:

- $\theta_k^{(t)}$ are updated with their weighted Maximum Likelihood Estimators for each k:

$$\begin{cases} \widehat{\xi}_k = \dfrac{1}{N_k} \sum_{i=1}^{n} r_{ik}^{(t-1)} \xi_i \\[2mm] \widehat{\psi}_k = \dfrac{1}{N_k} \sum_{i=1}^{n} r_{ik}^{(t-1)} \psi_i \\[2mm] \widehat{B}_k = N_k d \left(\sum_{i=1}^{n} r_{ik}^{(t-1)} \begin{pmatrix} \xi_i' - \widehat{\xi}_k' \\ \psi_i' - \widehat{\psi}_k' \end{pmatrix} (\Sigma_i^{-1}) \begin{pmatrix} \xi_i - \widehat{\xi}_k & \psi_i - \widehat{\psi}_k \end{pmatrix} \right)^{-1} \\[4mm] F_d\left(\dfrac{\widehat{\lambda}_k}{2}\right) = -\dfrac{1}{N_k} \sum_{i=1}^{n} r_{ik}^{(t-1)} \log\left(|\Sigma_i|\right) + d \log\left(\dfrac{N_k \widehat{\lambda}_k}{2}\right) - \log\left(\left|\sum_{i=1}^{n} r_{ik}^{(t-1)} \Sigma_i^{-1}\right|\right) \\[4mm] \widehat{\Lambda}_k = N_k \widehat{\lambda}_k \left(\sum_{i=1}^{n} r_{ik}^{(t-1)} \Sigma_i^{-1} \right)^{-1} \end{cases}$$

- $\pi_k^{(t)}$ are updated with N_k/n, $N_k = \sum_{i=1}^{n} r_{ik}$

4. **Repeat 2. and 3. until convergence**

Convergence is reached when the incomplete log-likelihood $l^{(t)}$ is unchanged between two consecutive iterations t and $t+1$ of the 2. and 3. steps:

$$l^{(t)} = \log\left(p(\{x_{1:N}\}|K, \{\theta_{1:K}^{(t)}\})\right) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k p(x_i|K, \{\theta_{1:K}^{(t)}\})\right)$$

G.2.2 MAP estimation via E-M algorithm

In order to avoid degenrate covariance matrices (for instance when K is set to too many clusters in the EM algorithm), it can be useful to replace MLE estimation with Maximum A Posteriori (MAP) estimations [Fraley and Raftery, 2007].

To perform a MAP estimation instead of a MLE estimation as in section G.2.1, the E-step of the algorithm is unchanged, but the M-step now maximizes the following Q function:

$$Q\left(\{\boldsymbol{\theta}_{1:K}\}\,\middle|\,\{\boldsymbol{\theta}_{1:K}^{(t-1)}\}\right) = \mathbb{E}\left[\log\left(p(\{\boldsymbol{\theta}_{1:K}\})p(x_{\{1:n\}},\ell_{\{1:n\}}|K,\{\boldsymbol{\theta}_{1:K}\}))\right)\,\middle|\,\{\boldsymbol{\theta}_{1:K}^{(t-1)}\}\right]$$

$$= \sum_{\ell_{\{1:N\}}}\left(\log\left(p(x_{\{1:n\}},\ell_{\{1:n\}}|K,\{\boldsymbol{\theta}_{1:K}^{(t-1)}\}))\right)\right) + \log(p(\{\boldsymbol{\theta}_{1:K}\}))$$

$$= \log(p(\{\boldsymbol{\theta}_{1:K}\})) + \sum_{k=1}^{K}\sum_{i=1}^{n} r_{ik}^{(t-1)}\log(\pi_k)$$

$$+ \sum_{k=1}^{K}\sum_{i=1}^{N} r_{ik}^{(t-1)}\log\left(p(x_i|K,\{\boldsymbol{\theta}_{1:K}\})\right)$$

We use the following priors :

— a Dirichlet prior over the cluster weigths $\pi_{\{1:K\}}$ with all parameters equal to the same α (if $\alpha = 1$, then this is equivalent to a uniform prior over the $K-1$ simplex):

$$(\pi_1,\ldots,\pi_K) \sim Dir(\alpha)$$

And for each k:

— a Normal-Wishart empirical bayes prior on $(\boldsymbol{\mu}_k, \boldsymbol{B}_k)$:

$$(\boldsymbol{\mu}_k, \boldsymbol{B}_k) \sim \mathcal{NW}(m, \kappa_0, \boldsymbol{C}, 4)$$

$$\boldsymbol{\mu}_k | m, \kappa_0, \boldsymbol{B}_k, \Sigma_{\{1:n\}} \sim \mathcal{N}\left(m, \frac{1}{\kappa_0}\left(\boldsymbol{B}_k \otimes \frac{1}{n}\sum_{i=1}^{n}\Sigma_i^{-1}\right)^{-1}\right)$$

$$\boldsymbol{B}_k | \boldsymbol{C} \sim \mathcal{W}(\boldsymbol{C}, 4)$$

with $m = \overline{\mu}_{\{1:n\}}$, $\boldsymbol{C} = 100\boldsymbol{I}_2$ and $\boldsymbol{L} = (\boldsymbol{S}^{(\xi)}+\boldsymbol{S}^{(\psi)})/2$ (where $\boldsymbol{S}^{(\xi)} = diag(var(\{\xi_{1:n}\}))$ and $\boldsymbol{S}^{(\psi)} = diag(var(\{\psi_{1:n}\}))$) and $\kappa_0 = 0.01$ for instance. The harmonic mean is used as an empirical bayes prior for the bloc variance matrix.

One can also specify a vague prior on $\boldsymbol{\mu}_k$: $\boldsymbol{\mu}_k \sim \mathcal{U}_{]-\infty,+\infty[}^{2d}$ (which simplifies the $\boldsymbol{\xi}$ and ψ MAP estimators, as long as no cluster has an exactly null 0 contribution N_k)

— a Wishart priors on Λ_k:

$$\Lambda_k \sim \mathcal{W}(\boldsymbol{L}, d+2)$$

with $\boldsymbol{L} = (\boldsymbol{S}^{(\xi)}+\boldsymbol{S}^{(\psi)})/2$ (where $\boldsymbol{S}^{(\xi)} = diag(var(\{\xi_{1:n}\}))$ and $\boldsymbol{S}^{(\psi)} = diag(var(\{\psi_{1:n}\}))$)

— an Exponential prior on λ_k under the constraint that $\lambda_k \geq d+1$:

$$\lambda_k - (d+1) \sim Exp(1)$$

The Q function is then:

$$Q(\theta|\theta^{(t-1)}) = \sum_{k=1}^{K} \left[-\frac{1}{2} \log \left(\left| B_k \otimes \left(\sum_{i=1}^{n} \Sigma_i^{-1} \right)^{-1} \right| \right) \right.$$

$$- \frac{\kappa_0}{2n}(\mu_k - m)' \left(B_k \otimes \sum_{i=1}^{n} \Sigma_i^{-1} \right) (\mu_k - m)$$

$$\left. + \frac{1}{2}\log(|B_k|) - \frac{1}{2}tr\left(C^{-1}B_k\right) + \frac{1}{2}\log(|\Lambda_k|) - \frac{1}{2}tr\left(L^{-1}\Lambda_k\right) - \lambda_k \right]$$

$$+ \sum_{k=1}^{K}\sum_{i=1}^{n} \left[r_{ik}^{(t-1)}\log(\pi_k) - \frac{\lambda_k + d + 1}{2}r_{ik}^{(t-1)}\log(|\Sigma_i|) \right.$$

$$- \frac{\lambda_k dr_{ik}^{(t-1)}}{2}\log(2) - r_{ik}^{(t-1)}\log(\Gamma_d(\frac{\lambda_k}{2})) + \frac{r_{ik}^{(t-1)}\lambda_k}{2}\log(|\Lambda_k|)$$

$$- \frac{r_{ik}^{(t-1)}}{2}\log(|B_k \otimes \Sigma_i|) - \frac{r_{ik}^{(t-1)}}{2}tr(\Lambda_k\Sigma_i^{-1})$$

$$\left. - \frac{r_{ik}^{(t-1)}}{2}(\mu_i - \mu_k)'(B_k \otimes \Sigma_i^{-1})(\mu_i - \mu_k) \right] + constant$$

and its partial derivatives yields:

$$\frac{dQ(\theta|\theta^{(t-1)})}{d\lambda_k} = -\frac{N_k}{2}\Gamma_d\left(\frac{\widehat{\lambda_k}}{2}\right) - \frac{1}{2}\sum_{i=1}^{n} r_{ik}^{(t-1)}\log\left(|\Sigma_i|\right) + \frac{N_k d}{2}\log\left(\frac{N_k\widehat{\lambda_k}}{2}\right)$$

$$- \frac{N_k}{2}\log\left(\left| \sum_{i=1}^{n} r_{ik}^{(t-1)}\Sigma_i^{-1} \right| \right) - 1$$

$$\frac{dQ(\theta|\theta^{(t-1)})}{dB_k} = \frac{d}{2}B_k + \frac{N_k d}{2}B_k - \sum_{i=1}^{n}\frac{r_{ik}^{(t-1)}}{2}\left(\begin{array}{c} \xi_i' - \xi_k' \\ \psi_i' - \psi_k' \end{array} \right)(\Sigma_i^{-1})\left(\xi_i - \xi_k \quad \psi_i - \psi_k \right)$$

$$- \frac{\kappa_0}{2n}\left(\begin{array}{c} \xi_k' - m^{(\xi)'} \\ \psi_k' - m^{(\psi)'} \end{array} \right)\sum_{i=1}^{n}\Sigma_i^{-1}\left(\xi_k - m^{(\xi)} \quad \psi_k - m^{(\psi)} \right) + \frac{1}{2}B_k - \frac{1}{2}C^{-1}$$

$$\frac{dQ(\theta|\theta^{(t-1)})}{d\Lambda_k} = \frac{N_k\lambda_k}{2}\Lambda_k^{-1} - \frac{1}{2}\sum_{i=1}^{n} r_{ik}^{(t-1)}\Sigma_i^{-1} + \frac{1}{2}\Lambda_k^{-1} - \frac{1}{2}L^{-1}$$

The MAP estimators of $\theta_k | \theta_k^{(t-1)}$ are thus:

$$
\begin{cases}
\widehat{\pi}_k^{MAP} = \dfrac{N_k + \alpha - 1}{n + K(\alpha - 1)} \\[4mm]
\widehat{\xi}_k^{MAP} = \dfrac{\kappa_0 m^{(\xi)}/n + \sum_{i=1}^{n} r_{ik}^{(t-1)} \xi_i}{\kappa_0/n + N_k} \\[4mm]
\widehat{\psi}_k^{MAP} = \dfrac{\kappa_0 m^{(\psi)}/n + \sum_{i=1}^{n} r_{ik}^{(t-1)} \psi_i}{\kappa_0/n + N_k} \\[4mm]
\widehat{B}_k^{MAP} = (N_k d + d + 1) \Bigg[C^{-1} \\[4mm]
\qquad + \sum_{i=1}^{n} r_{ik}^{(t-1)} \begin{pmatrix} \xi_i' - \widehat{\xi}_k^{MAP}{}' \\ \psi_i' - \widehat{\psi}_k^{MAP}{}' \end{pmatrix} \left(\Sigma_i^{-1} \right) \begin{pmatrix} \xi_i - \widehat{\xi}_k^{MAP} & \psi_i - \widehat{\psi}_k^{MAP} \end{pmatrix} \\[4mm]
\qquad + \dfrac{\kappa_0}{n} \begin{pmatrix} \xi_k' - m^{(\xi)}{}' \\ \psi_k' - m^{(\psi)}{}' \end{pmatrix} \sum_{i=1}^{n} \Sigma_i^{-1} \begin{pmatrix} \xi_k - m^{(\xi)} & \psi_k - m^{(\psi)} \end{pmatrix} \Bigg]^{-1} \\[6mm]
0 = N_k F_d \left(\dfrac{\widehat{\lambda}_k^{MAP}}{2} \right) + \sum_{i=1}^{n} r_{ik}^{(t-1)} \log \left(|\Sigma_i| \right) - N_k d \log \left(\dfrac{N_k \widehat{\lambda}_k^{MAP}}{2} \right) \\[4mm]
\qquad + N_k \log \left(\left| \sum_{i=1}^{n} r_{ik} \Sigma_i^{-1} \right| \right) + 2 \\[6mm]
\widehat{\Lambda}_k^{MAP} = \left(N_k \widehat{\lambda}_k^{MAP} + 1 \right) \left(L^{-1} + \sum_{i=1}^{n} r_{ik}^{(t-1)} \Sigma_i^{-1} \right)^{-1}
\end{cases}
$$

with $N_k = \sum_{i=1}^{n} r_{ik}^{(t-1)}$.

The corresponding E-M algorithm for MAP estimation can therefore be written as follows:

1. **Initialization**

   $\theta_k^{(0)}$ are initialized randomly (π_k are initialized at $1/K$)

2. **E step**

   Compute the membership weights $r_{ik}^{(t-1)}$ for each observation $i = 1 \ldots N$ for each cluster $k = 1 \ldots K$:

3. M step

Update the parameters:

- θ_k are updated with their MAP estimation for each k:

$$
\begin{cases}
\pi_k^{(t)} = \dfrac{N_k + \alpha - 1}{n + K(\alpha - 1)} \\[3mm]
\xi_k^{(t)} = \dfrac{\kappa_0 m^{(\xi)}/n + \sum_{i=1}^n r_{ik}^{(t-1)} \xi_i}{\kappa_0/n + N_k} \\[3mm]
\psi_k^{(t)} = \dfrac{\kappa_0 m^{(\psi)}/n + \sum_{i=1}^n r_{ik}^{(t-1)} \psi_i}{\kappa_0/n + N_k} \\[3mm]
B_k^{(t)} = (N_k d + d + 1)\left[C^{-1} + \sum_{i=1}^n r_{ik}^{(t-1)} \begin{pmatrix} \xi_i' - \xi_0' \\ \psi_i' - \psi_0' \end{pmatrix} (\Sigma_i^{-1}) \begin{pmatrix} \xi_i - \xi_0 & \psi_i - \psi_0 \end{pmatrix} \right. \\[3mm]
\qquad\qquad \left. + \dfrac{\kappa_0}{n} \begin{pmatrix} \xi_k^{(t)\,\prime} - m^{(\xi)\,\prime} \\ \psi_k^{(t)\,\prime} - m^{(\psi)\,\prime} \end{pmatrix} \sum_{i=1}^n \Sigma_i^{-1} \begin{pmatrix} \xi_k^{(t)} - m^{(\xi)} & \psi_k^{(t)} - m^{(\psi)} \end{pmatrix} \right]^{-1} \\[3mm]
F_d\left(\dfrac{\lambda_k^{(t)}}{2} \right) = -\dfrac{1}{N_k} \sum_{i=1}^n r_{ik}^{(t-1)} \log\left(|\Sigma_i|\right) + d\log\left(\dfrac{N_k \lambda_k^{(t)}}{2} \right) - \log\left(\left| \sum_{i=1}^n r_{ik}^{(t-1)} \Sigma_i^{-1} \right| \right) \\[3mm]
\Lambda_0^{(t)} = (N_k \lambda_k^{(t)} + 1)\left(L^{-1} + \sum_{i=1}^n r_{ik}^{(t-1)} \Sigma_i^{-1} \right)^{-1}
\end{cases}
$$

- $\pi_k^{(t)}$ are updated with N_k/n, $N_k = \sum_{i=1}^n r_{ik}$

4. Repeat 2. and 3. until convergence

Convergence is reached when the incomplete log-likelihood $l^{(t)}$ is unchanged between two consecutive iterations t and $t + 1$ of the 2. and 3. steps:

$$
l^{(t)} = \log\left(p(\{x_{1:N}\}|K, \{\theta_{1:K}^{(t)}\}) \right) = \sum_{i=1}^n \log\left(\sum_{k=1}^K \pi_k p(x_i|K, \{\theta_{1:K}^{(t)}\}) \right)
$$

Résumé
Analyse intégrative de données de grande dimension
appliquée à la recherche vaccinale

Les données d'expression génique sont reconnues comme étant de grande dimension, et nécessitant l'emploi de méthodes statistiques adaptées. Mais dans le contexte des essais vaccinaux, d'autres mesures, comme par exemple les mesures de cytométrie en flux, sont également de grande dimension. De plus, ces données sont souvent mesurées de manière longitudinale. Ce travail est bâti sur l'idée que l'utilisation d'un maximum d'information disponible, en modélisant les connaissances a priori ainsi qu'en intégrant l'ensemble des différentes données disponibles, améliore l'inférence et l'interprétabilité des résultats d'analyses statistiques en grande dimension. Tout d'abord, nous présentons une méthode d'analyse par groupe de gènes pour des données d'expression génique longitudinales. Ensuite, nous décrivons deux analyses intégratives dans deux études vaccinales. La première met en évidence une sous-expression des voies biologiques d'inflammation chez les patients ayant un rebond viral moins élevé à la suite d'un vaccin thérapeutique contre le VIH. La deuxième étude identifie un groupe de gènes lié au métabolisme lipidique dont l'impact sur la réponse à un vaccin contre la grippe semble régulé par la testostérone, et donc lié au sexe. Enfin, nous introduisons un nouveau modèle de mélange de distributions t asymétriques à processus de Dirichlet pour l'identification de populations cellulaires à partir de données de cytométrie en flux disponible notamment dans les essais vaccinaux. En outre, nous proposons une stratégie d'approximation séquentielle de la partition a posteriori dans le cas de mesures répétées. Ainsi, la reconnaissance automatique des populations cellulaires pourrait permettre à la fois une avancée pratique pour le quotidien des immunologistes ainsi qu'une interprétation plus précise des résultats d'expression génique après la prise en compte de l'ensemble des populations cellulaires.

Mots clés : Analyse intégrée ; Analyse par groupe de gènes ; Bayésien non paramétrique ; Connaissance a priori ; Cytométrie en flux ; Dimorphisme sexuel ; Distribution t asymétrique ; Données de grande dimension ; Fenêtrage automatisé ; Grippe ; Génomique ; Modèle de mélange ; Processus de Dirichlet ; Vaccin ; VIH.

Abstract
Integrative analysis of high-dimensionnal data
applied to vaccine research

Gene expression data is recognized as high-dimensional data that needs specific statistical tools for its analysis. But in the context of vaccine trials, other measures, such as flow-cytometry measurements are also high-dimensional. In addition, such measurements are often repeated over time. This work is built on the idea that using the maximum of available information, by modeling prior knowledge and integrating all data at hand, will improve the inference and the interpretation of biological results from high-dimensional data. First, we present an original methodological development, Time-course Gene Set Analysis (TcGSA), for the analysis of longitudinal gene expression data, taking into account prior biological knowledge in the form of predefined gene sets. Second, we describe two integrative analyses of two different vaccine studies. The first study reveals lower expression of inflammatory pathways consistently associated with lower viral rebound following a HIV therapeutic vaccine. The second study highlights the role of a testosterone mediated group of genes linked to lipid metabolism in sex differences in immunological response to a flu vaccine. Finally, we introduce a new model-based clustering approach for the automated treatment of cell populations from flow-cytometry data, namely a Dirichlet process mixture of skew t-distributions, with a sequential posterior approximation strategy for dealing with repeated measurements. Hence, the automatic recognition of the cell populations could allow a practical improvement of the daily work of immunologists as well as a better interpretation of gene expression data after taking into account the frequency of all cell populations.

Key words: Automated gating; Dirichlet process; Flow cytometry; Gene set analysis; High-dimensional data; HIV; Influenza; Integrative analysis; Mixture model; Nonparametric Bayesian; Prior knowledge; Sexual dimorphism; Skew t-distribution; Statistical genomics; Vaccine.