Universidad Autónoma de Madrid

FACULTAD DE MEDICINA

Departamento de Medicina Preventiva y Salud Pública

Université de Liège

SCIENCES APPLIQUÉES

Département d'Electricité, Electronique et Informatique

PhD Thesis – Statistical Genetics

STATISTICAL METHODS FOR THE INTEGRATION ANALYSIS OF –OMICS DATA (GENOMICS, EPIGENOMICS AND TRANSCRIPTOMICS): AN APPLICATION TO BLADDER CANCER

Author

Silvia Pineda San Juan

Supervisors

Núria Malats

Spanish National Cancer Research Centre

Genetic and Molecular Epidemiology

Madrid, Spain

Kristel Van Steen

Systems and Modeling Unit

Montefiore Institute

Liége, Belgium

Madrid, Spain, September 15th, 2015

Panel Defense

Fernando Rodríguez Artalejo, Ph.D.
(Department of Preventive Medicine and Public Health, UAM, Spain)

Alfonso Valencia, Ph.D.
(Structural Biology and Biocomputing programme, CNIO, Spain)

Douglas Easton, Ph.D.
(Centre for Cancer Genetic Epidemiology, Department of Public Health, Cambridge, UK)

Monika Stoll, Ph.D.
(Department of Genetic Epidemiology, University of Munster, Germany)

Mario F. Fraga, Ph.D.
(Asturias Central University Hospital, CSIC, Spain)

Fátima Al-Shahrour, Ph.D. (Substitute)
(Translational Bioinformatics unit, CNIO, Spain)

Stephan Ossowski, Ph.D. (Substitute)
(Genomic and Epigenomic Variation Disease, CRG, Spain)

To my sister María,

To my parents,

To Mikel,

Acknowledgements

In the first place, I would like to thank to my supervisors, Núria Malats and Kristel Van Steen, for their support during my entire PhD study, for her patience, motivation, enthusiasm, the immense scientific knowledge and also personal advice that will be all very important in the development of my career. I thank Núria for giving me the opportunity to work with her and her group. She trusted me from the beginning and she always encouraged me with new challenges to make me think and go deeper into each scientific issue. I thank Kristel to agree on this collaboration and received me in her group for an entire year. I am very happy to have met and work with both of them and I am very grateful for the opportunity of this joint PhD opportunity.

I also want to give a special thanks to Roger Milne. He guided me two first years I was in the CNIO when I was completely lost in this new thing it was for me "research". I learnt many things from him scientific and personally.

I also want to thank the members of my CNIO committee thesis, Alfonso Valencia, Manuel Hidalgo and Peter Van Der Spek for their comments and suggestions during my PhD that made me progress on my research.

I would like to acknowledge the principal investigators, monitors and patients from the Spanish Bladder Cancer (SBC)/EPICURO study that have generated all the data used in this thesis, and Francisco Real who has financed the omics data generation and has contributed in many discussions of this thesis.

I also want to thank Fernando Rodríguez Artalejo, the director of my doctoral program in the UAM, for all the general talks we have had about science that I really enjoyed. Also to him and to Axelle Lambotte, from the administrative office in the ULg, for all their help with the administrative issues from both universities that have been sometimes very complicated and frustrated.

I also want to thank all the funding support that have made possible this thesis, la Obra Social Fundación La Caixa, the Short-Term Scientific Mission (STSM) from COST Action BM1204, and the ULg fellowship for foreign students.

I would like to thank to all the members of the Molecular and Genetic Epidemiology group in CNIO, to Roger, Toni, Gäelle, Salman, Mat, Raquel, Jesús, André, Alexandra, Marien, Marina that have already left the group, to Evangelina and Mirari that are here since I started. To Paulina, Marta, Esther and Veronica that have been the last incorporations to the group. I really want to thank to all of them for all the discussions we have had in the group, the talks during the lunch, the Friday beers and also the life outside the CNIO doors. I would like to give a special thanks to Toni for his friendship and all the geeks talks that I miss them since he left, to Paulina for all her help during my writing, for reading my thesis and for her valuable comments, and mainly for her patience in the most critical moments during this process and her friendship inside and outside the CNIO. I do not want to forget Ángel from the bioinformatics unit in CNIO who has helped me every time I was lost in this Linux world (I still owe him some beers) and Guille from CEGEN unit in CNIO, who has helped me in understanding and solving many of the genetic issues.

I also thank to the members of the Systems and Modeling Unit from the Montefiore Institute in the ULg, Kiril, Ramouna, Elena, Françoise and Kris for welcome me in the group and explained and shared with me their projects. I would like to give a special thanks to Kiril, Ramouna and Kris for all the time we have passed together during the entire year I was in Liège. I really had a good time going to all the restaurants to taste the food from all of our countries and all the interesting talks we had trying to explain our different cultures to each other.

Because having fun is also an important part of doing a PhD, I would like to give a special thanks to all my friends, for all the good times we have passed together.

Especialmente, quiero agradecer a mis padres que me han enseñado los valores importantes de la vida, a mi hermana de la que admiro la disciplina que tiene en el deporte y que intento aprender cada día. Son las personas más importantes en mi vida, nunca dudaron que este era el camino que debía escoger y me apoyaron en cada decisión que fui tomando confiando en mí y simplemente estando en cada momento para cualquier cosa. También quiero agradecer tremendamente a Mikel, mi compañero, ya que me ha apoyado y acompañado durante todo este proceso y en la parte más emocional de estudiar un doctorado, especialmente estos últimos meses que han sido los más complicados. Siempre me convencía de que yo podía hacerlo y no solo me ha apoyado en esta etapa, sino que juntos empezamos la próxima etapa de la que seguro aprenderemos, nos reiremos, tendremos buenos momentos y los malos momentos quedarán para las historietas.

At last, to all the people that has crossed in my way during these last four years because have contributed in somehow to this thesis.

Summary

An increase amount of –omics data is being generated and single –omics analyses have been performed to analyze it in the last decades. While the effort has revealed significant findings to better understand the biology of complex disease, such as cancer, combining more than two –omics data will certainly explain further biological insights not found otherwise. For this reason, in the last five years the idea of integrating data has appeared in the context of system biology. However, the integration of –omics data requires of appropriate statistical techniques to address the main challenges that high-throughput data impose. In this thesis, we propose different statistical approaches to integrate –omics data (genomics, epigenomics, and transcriptomics from tumor tissue, and genomics from blood samples) in individuals with bladder cancer. In the first contribution, a framework based on a multi-staged strategy was proposed. Pairwise combinations using the three –omics measured in tumor were analyzed (transcriptomics-epigenomics, eQTL and methQTL) to end with the combination of all of them in triples relationships. The results showed a whole spectrum of relationships and sound biological *trans* associations identifying new possible molecular targets. In the second contribution, a multi-dimensional analysis was applied to the three –omics considered together in the same model. Penalized regression methods (LASSO and ENET) were applied since they can combine the data in a large input matrix dealing with many of the –omics data integration challenges. Besides, a permutation–based MaxT method was proposed to assess the goodness of fit while correcting for multiple testing which are the main drawbacks of the penalized regression methods. We obtained a list of genes associated with genotypes and DNA methylation in *cis* relationship that were further externally validated in an independent data set. Finally, the same approach was applied to integrate the three –omics data in tumor with the genomics data in blood samples in an integrative eQTL analysis. This approach was compared with the 2 stage regression (2SR) approach previously used for eQTL integrative analysis. Our approach highlighted relevant eQTLs including the ones found by the 2SR strategy generating a list of genes and eQTLs that may be considered in future analyses. Overall, we have shown that –omics integrative analysis is needed to find missing or hidden information. To this end, applying appropriate statistical approaches is needed identify sound biological relationships.

Resumen

En las últimas décadas, la cantidad de datos -ómicos generados ha incrementado considerablemente y con ellos, se han realizado múltiples análisis considerando cada dato –ómico por separado. Este tipo de análisis ha revelado hallazgos significativos para entender mejor las enfermedades complejas, como el cáncer, pero la combinación de más de dos conjuntos de datos –ómicos puede revelar nuevos conocimientos biológicos que no se podrían encontrar de otra forma. Así, en los últimos cinco años, ha aparecido el concepto de integración de datos en el contexto de la biología de sistemas. No obstante, la integración de datos –ómicos requiere de técnicas estadísticas apropiadas para hacer frente a los principales retos que los datos de alto rendimiento (-ómicos) imponen. En esta tesis, proponemos diferentes aproximaciones estadísticas para integrar datos –ómicos (genómica, transcriptómica y epigenómica del tejido tumoral y genómica de sangre) en individuos con cáncer de vejiga. Como un primer enfoque, se propone un marco basado en una estrategia de etapas múltiples donde se analizan todas las posibles combinaciones por parejas utilizando los tres datos -ómicos medidos en el tejido tumoral (transcriptómica-epigenómica, eQTL y methQTL) para finalmente, combinar los resultados significativos en relaciones triples. Estas relaciones sugieren patrones y asociaciones biológicas *trans* muy interesantes. Como segundo enfoque, se propone un análisis multi-dimensional, donde los tres datos -ómicos se consideran conjuntamente en el mismo modelo. Para ello, se han aplicado métodos de regresión penalizada (LASSO y ENET) ya que pueden combinar los datos en una misma matriz de entrada haciendo frente a muchos de los retos que la integración de datos –ómicos impone. Además se propone un método basado en permutaciones MaxT para evaluar la bondad de ajuste a la vez que se corrige por test múltiples ya que precisamente estos representan los dos inconvenientes principales de los métodos de regresión penalizada. Como resultado una lista de genes asociados con genotipos y metilación del ADN en relaciones *cis* que ha sido validada en una base de datos externa. Por último, este mismo enfoque se ha implementado para integrar los tres datos -ómicos en tumor con la genómica en las muestras de sangre en un análisis de integración de eQTLs y se ha comparado con una regresión en 2 etapas ya que es un método previamente utilizado para el análisis de integración de eQTLs. Nuestro enfoque muestra relevantes eQTLs además de las ya propuestas por la regresión en 2 etapas generando una lista de genes y eQTLs que pueden ser consideradas en análisis futuros. En general, esta tesis muestra lo necesario que son los análisis de integración de varios datos –ómicos para encontrar información que todavía no conocemos. Además demostramos que la implementación de métodos estadísticos apropiados es imprescindible identificar relaciones biológicas robustas.

Résumé

Ces dernières décennies, la quantité de données -omiques générées a considérablement augmenté résultant en de multiples analyses des données -omiques considérés séparément. Ce type d'analyse a permis des avancées significatives dans la comprehension des maladies complexes comme le cancer, par conséquent la combinaison de plusieurs ensembles de données –omiques entre elles pourrait permettre d'approfondir encore les connaissances biologiques. Ainsi, ces cinq dernières années, est apparu le concept d'intégration des données dans le contexte de la biologie des systèmes. Cependant, l'intégration des données -omiques exige l'application de méthodes statistiques appropriées pour relever les défis majeurs imposés par les données de haute performance (-omiques). Dans cette thèse, nous proposons différentes approches statistiques pour intégrer données -omiques (génomique, transcriptomique et epigenomique au sein de tissu tumoral et génomique des échantillons de sang) chez des personnes atteintes d'un cancer de la vessie. Une première approche est fondée sur une stratégie en plusieurs étapes. Toutes les combinaisons possibles de paires ont été analysées en utilisant les trois données -omiques mesurées dans le tissu tumoral (transcriptomique-épigénomique, eQTL et methQTL) pour terminer avec la combinaison de chacun d'eux dans les triples relations. Nous avons montré un large spectre d'associations entre elles et les associations *trans-acting* fiables qui ont permis d'identifier de nouvelles cibles moléculaires potentielles. Une deuxième approche consiste en une analyse multi-dimensionnelle, où les trois données -omiques étaient considérées ensemble dans le même modèle. À cette fin, des méthodes de régressions pénalisées ont été appliquées (LASSO et ENET), permettant de relever les defis de l'integration de données -omiques en les entrant dans de larges matrices. Les permutations MaxT ont permis d'évaluer la qualité de l'ajustement tout en corrigeant pour les tests multiples qui sont les principaux inconvénients des méthodes de régressions pénalisées. Nous avons identifié et validé dans une base de données externe une liste de gènes associés aux génotypes et à la méthylation de l'ADN dans les relations *cis*. Cette même approche a été appliquée pour intégrer les trois -omiques tumorales et les données génomiques des échantillons de sang en une analyse d'intégration des eQTL. Cette méthode a été comparée avec la régression en 2 étapes déjà utilisée pour l'intégration des eQTLs. Notre approche a mis en lumière des eQTLs d'intérêt comprenant ceux déjà proposés par la régression en 2 étapes et permettant de générer une liste de gènes et d'eQTLs qui pourront être prises en compte dans les analyses futures. Au total, cette thèse a montré que l'intégration des données -omiques est nécessaire pour l'identification d'informations manquantes, cachées ou fausses. L'application de méthodes statistiques appropriées est indispensable pour identifier des relations biologiques solides.

Contents

List of Figures

PART 1 – Chapter 1

PART 1 – Chapter 2

PART 1 – Chapter 4

PART 1 – Chapter 5

PART 2 – Chapter 2

PART 2 – Chapter 2

PART 2 – Chapter 3

PART 3 – Chapter 1

PART 3 – Chapter 2

PART 3 – Chapter 3

List of Tables

PART 1 – Chapter 5

PART 2 – Chapter 1

PART 2 – Chapter 2

PART 3 – Chapter 1

PART 3 – Chapter 2

PART 3 – Chapter 3

Abbreviations

SBC/EPICURO	Spanish Bladder Cancer/EPICURO
TCGA	The Cancer Genome Atlas study
DNA	Deoxyribonucleic acid
SNP	Single Nucleotide Polymorphisms
CpG	Cytosine-phosphate-Guanine
mRNA	Messenger RNA
GWAS	Genome Wide Association Studies
eQTL	Expression Quantitative Trait Loci
methQTL	Methylation Quantitative Trait Loci
OLS	Ordinary Least Square
PCA	Principal Component Analysis
FA	Factor Analysis
LD	Linkage Disequilibrium
PCs	Principal Components
CCA	Canonical Correlation Analysis
MFA	Multiple Factor Analysis
SCCA	Sparse Canonical Correlation Analysis
PLS	Partial Least Squares
MSE	Mean Squared Error
LASSO	Least Absolute Shrinkage and Selection Operator
ENET	Elastic Net
CV	Cross Validation
ASR	Age Standardize Rate
UBC	Urothelial Bladder cell Carcinoma
NMI	Non-muscle Invasive
MI	Muscle Invasive
TCGA	The Cancer Genome Atlas
QC	Quality Control
MAF	Minor Allele Frequency
RMA	Robust Multi-Array Average

LOI	Loss Of Imprinting
2SR	2-Stage Regression
MLR	Multiple Linear Regression

PART 1.

INTRODUCTION AND AIMS

One important goal in human genetics and molecular epidemiology is to elucidate the genetic architecture of complex diseases. Important technological advances are crucial to better characterize the different layers of the biological processes that are involved in a complex disease. However, the vast amount of molecular –omics data that are generated (genomics, epigenomics, transcriptomics, proteomics, metabolomics, among others) needs from optimal modelling to extract the most information possible that has been hidden or missing until now. Data integration is the technique to process the different types of –omics data as combinations of predictor variables to allow comprehensive modelling of complex diseases or phenotypes. The frame of the work presented here is under the umbrella of data integration focusing on the methodological aspects of the process to perform an integrative –omics analysis.

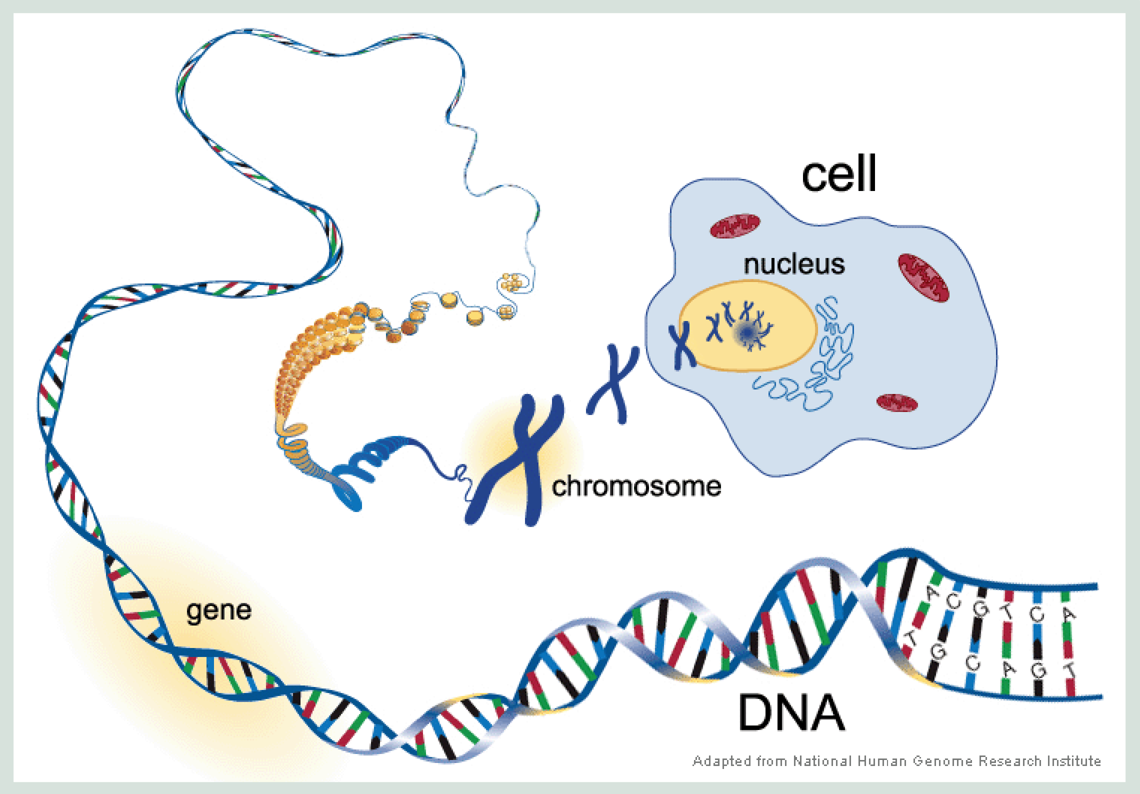
In the introduction of this thesis a description of the –omics data used for this work is discussed from the simplest to the most complex relationships in terms of molecular description, bioinformatics and statistical methods providing biological information and examples of studies that have been performed in the context of data integration. In Chapter 1, the three types of data used in the thesis (genomics, epigenomics and transcriptomics) are described. In chapter 2, the biological combinations of these three types of data and the main statistical methods used to analyze them are described. In chapter 3, the overview and challenges of integrative –omics analysis and the main statistical methods for data integration are described and in chapter 4 the penalized regression methods used in this thesis to perform the data integration are detailed described. In chapter 5 is explained the epidemiology, tumorigenesis and etiology of bladder cancer and the studies used in this thesis (the Spanish Bladder Cancer (SBC)/EPICURO study and The Cancer Genome Atlas study (TCGA)) and chapter 6 provides the hypothesis, objectives and organization of the thesis.

Chapter 1: Introduction to –omics data

This chapter introduces the three –omics data (genomics, epigenomics and transcriptomics) used in this thesis. A short description of the concept of each dataset and their functions in the human biological system is provided.

1.1. Genomics

Genomics is considered as the study of the genomic DNA (deoxyribonucleic acid) data that is available in many species. The human genome is the complete sequence of the genetic information of humans and is stored in each cell. Within cells, DNA is packed in the nucleus and in the mitochondrias. Here I will only refer to the first DNA. Genetic information is organized into structures called chromosomes and is encoded as the DNA molecule. The DNA consists of two strands containing millions of nucleotides. The nucleotides are organic molecules that serves as a subunits and are composed of a nitrogen nucleobase (guanine (G), adenine (A), thymine (T) and cytosine (C)), a five-carbon sugar and at least one phosphate group. This information includes protein-coding genes and non-coding sequences (Figure 1.1.1).



Adapted from National Human Genome Research Institute

Figure 1.1.1 Location and structure of the DNA molecule in the human genome. (Copied from National Human Genome Research Institute (https://www.genome.gov/))

The Human Genome Project produced the first complete DNA sequence of individual human genomes in 2001 (Venter et al. 2001) with a consensus of approximately three billion nucleotide positions. DNA sequencing is the process to determine the order of all the nucleotides within the DNA molecule. It is now possible to collect the whole genetic information from each individual in a study using whole genome sequencing and this will be a very important achievement in the future of the personalized medicine. Most of the studies until now, including this thesis, have determined a subset of genetic markers to capture as much of the complete genome information as possible. The markers used are Single Nucleotide Polymorphisms (SNPs) that are changes of one nucleotide base pair that occurs in at least 1% of the population. In humans, the majority of the SNPs are bi-allelic, indicating the two possible bases at the corresponding position within a gene. If we define *A* as the common allele and *B* as the variant allele, three combinations are possible: *AA* (the common homozygous), *AB* (the heterozygous) and *BB* (the variant homozygous). These combinations are known as the genotypes and they are assessed with SNP genotyping platforms.

1.2. Epigenomics

Epigenomics is the study of all the epigenetic modifications that occur on the genetic material in a cell without alterations in the DNA sequence. These changes mainly include DNA methylation and histones modifications. DNA methylation is associated with a number of very important processes (genomic imprinting, X-chromosome inactivation, suppression of repetitive elements, and regulation of cell specific gene expression) (Bird 2002), being the most studied epigenetic marker. In humans, DNA methylation involves the addition of a methyl group to the 5’ position of the cytosine at a Cytosine-phosphate-Guanine (CpG) dinucleotide by DNA methyltransferase (DNMT) enzymes (Figure 1.1.2).

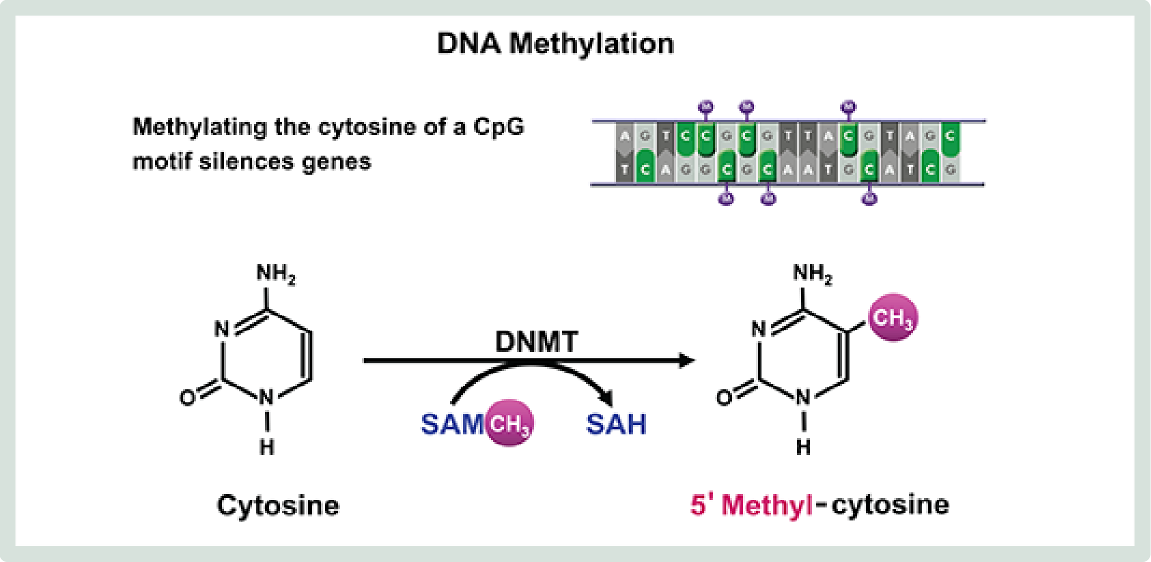


Figure 1.1.2. Representation of DNA methylation with the addition of a methyl group (-CH_3) to the 5' position of the cytosine. (Copied from Samir Zakhari. The NIAAA journal 2013;35 (1):6-16)

They are distributed over the human genome with the exception of some regions with high density of CpG dinucleotides that are denominated CpG islands. These specific regions are often located in gene promoters (the region that facilitates transcription of a particular gene) and they are usually unmethylated in normal cells. When methylated, often are associated with gene silencing. The CpG shores, located at 2kb from the island's boundaries are also important in gene regulation. Alterations in DNA methylation may affect phenotypic transmission and may be part of the etiology of human disease (Robertson & Wolffe 2000; Portela & Esteller 2010) and are very well implicated in carcinogenesis (Esteller 2008). To assess information of the CpG sites in the genome the methylation beadchip platforms are used.

1.3. Transcriptomics

Transcriptomics is the study of the complete set of RNA transcripts that are produced by the genome. This process is called transcription and it is the first step of gene expression in which a particular segment of DNA is copied into RNA by the enzyme RNA polymerase. In the process of transcription the two DNA may be labeled as antisense strand that serves for the production of the RNA transcript and the sense strand which includes the DNA version of the transcript sequence. The antisense strand is identical to the sense strand with the exception that thymines (T) are replaced with uraciles (U) in the RNA (Figure 1.1.3).

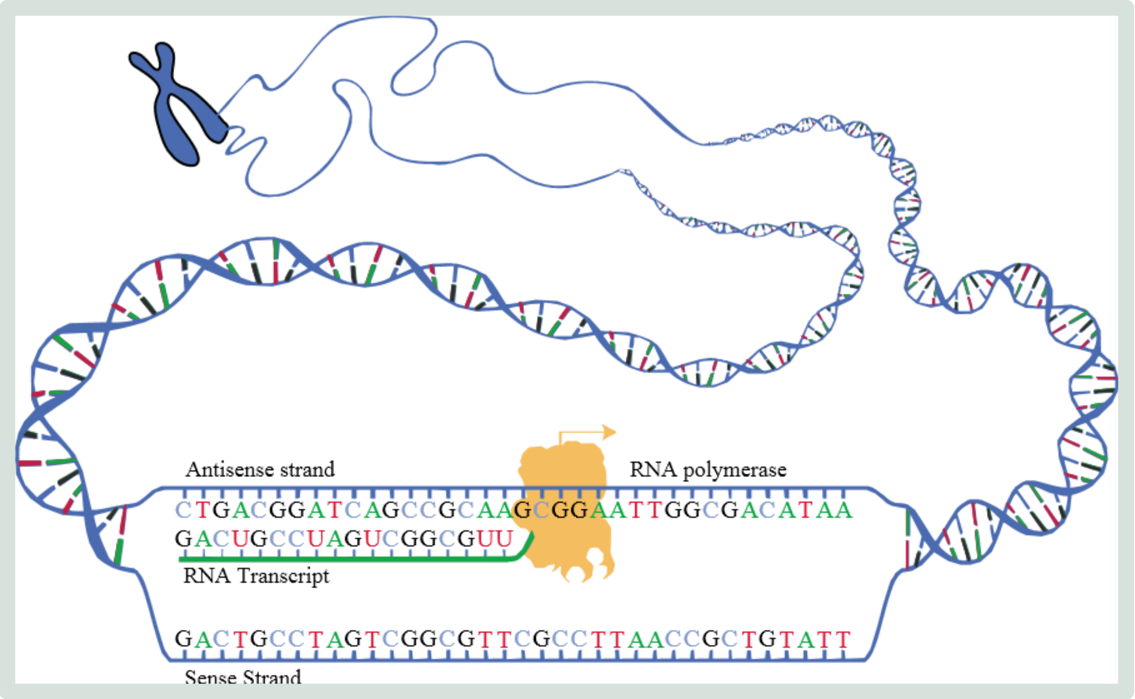


Figure 1.1.3. Synthesis of mRNA copied from the DNA base sequences by RNA polymerase. (Copied from Bioknowledgy webpage)

The RNA molecule encodes at least one gene that will be transcribed as messenger RNA (mRNA) if the gene transcribed encodes a protein, or non-coding RNA (microRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), or other enzymatic RNA molecules if not. mRNA abundance can be used as an indirect measure of gene expression. The expression of different genes allows cells to differentiate and perform different functions. There are an estimated 20,000-25,000 human protein-coding genes (Human Genome Sequencing Consortium International 2004; Pennisi 2012) whose mRNA transcript levels can be measured using high-throughput data with microarray platforms.

Chapter 2. From one –omics analysis to pairwise –omics analysis

The –omics data shown in Chapter 1 have been extensively studied to better understand and characterize complex diseases, such as cancer. The individual analysis of each –omics data has provided an interesting amount of new findings in the last decades. Genomics have been mainly studied through SNPs in Genome Wide Association Studies (GWAS). Transcriptomics have been also extensively studied with differential gene expression analysis, and epigenomics, a less studied –omics data type, has been also assessed through Epigenomics Wide Association Studies (EWAS) (Rakyan et al. 2011) following the success of GWAS. In this chapter, I give an overview of the analysis of the pairwise combinations between these three types of data and the main statistical methods to analyze them.

2.1. Epigenomics – Transcriptomics

In 1975, it was first suggested that DNA methylation was involved in gene regulation (Riggs 1975) showing the X chromosome inactivation process. Since then, the study of the relationship between the CpG sites and the gene silencing became very important. Also, the position of the CpG sites in the genome, and especially with respect to the gene, influences this relationship. For example, methylated CpG islands and shores located in promoter regions of the gene may act blocking the expression of the gene while if they are located within the gene body, it might stimulate transcription (Jones 2012) (Figure 1.2.1).

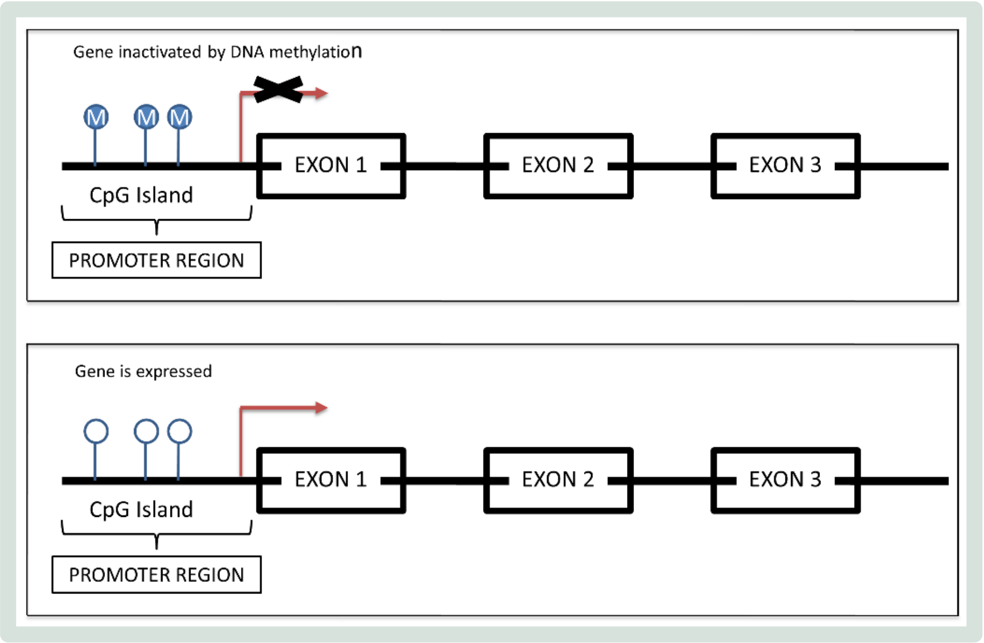


Figure 1.2.1. Inactivation of a gene by DNA methylation

The relationship between DNA methylation and gene expression is very important to better understand the complexity of the traits. In this regard, hypermethylation of CpG islands and shores in the promoter region of a tumor-suppressor gene is a major event in many cancers (Portela & Esteller 2010). Many studies have been performed looking at relationships between these two –omics data types.

2.2. Genomics – Transcriptomics

The expression levels of many genes shows abundant natural variation and this variation of many genes has a heritable component in humans (Morley et al. 2004). Studies usually assess whether gene expression levels, measured as a quantitative phenotype, are significantly associated with genetic variation (SNP genotyped). This association is known as expression Quantitative Trait Loci (eQTL) and it has been extensively studied (Stranger et al. 2007; Zhernakova et al. 2013; Cheung & Spielman 2009; Bryois et al. 2014), also linked with diseases (Nica et al. 2010; Nicolae et al. 2010; Westra et al. 2013; Shpak et al. 2014). Normally, they are categorized according to the distance between the SNP and the target gene. The last agreement for this definition refers to *cis*-acting eQTL if the distance is within 1MB window of the gene (the SNP is located within 1MB upstream and 1MB downstream the gene) and *trans*-acting, otherwise (Figure 1.2.2).

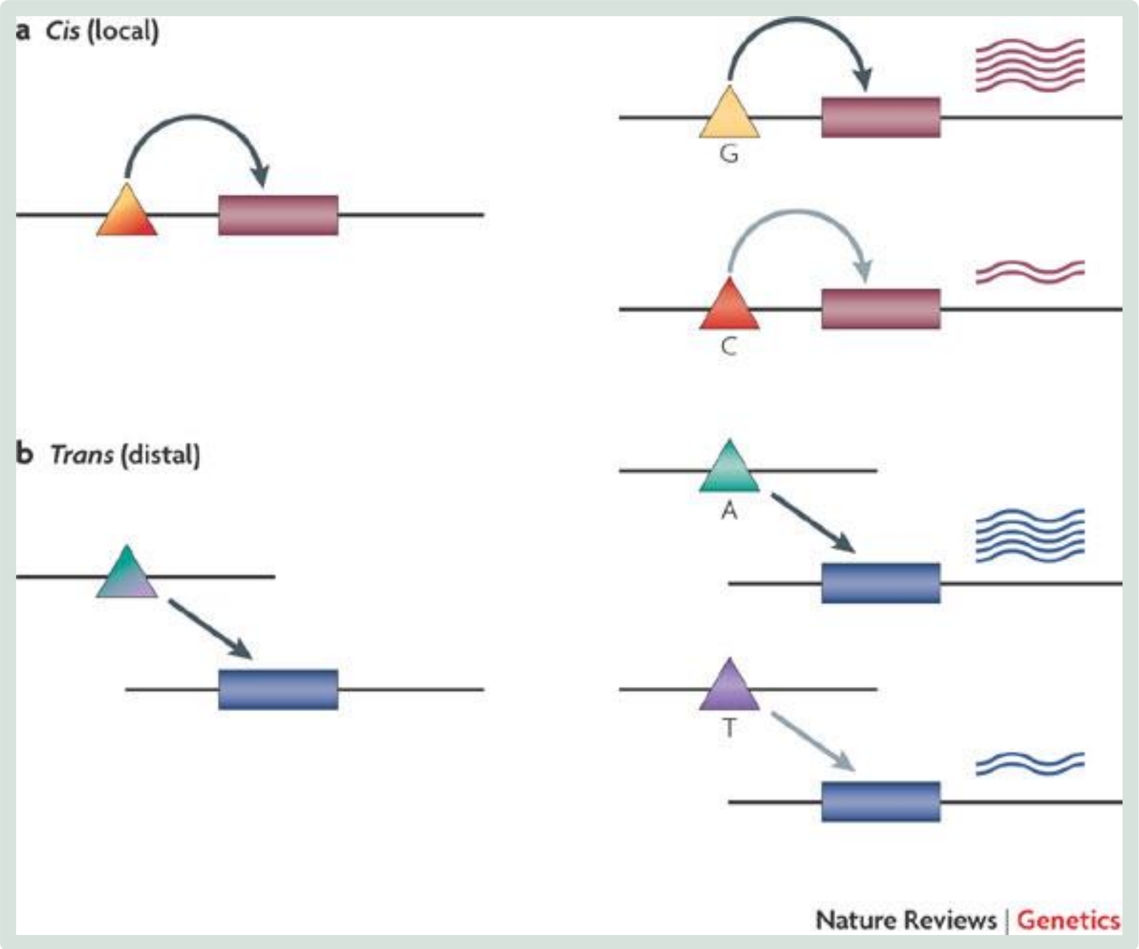


Figure 1.2.2. SNP regulating in *cis* (a) and *trans* (b) the expression of a gene. *Cis*-acting is close to the target gene while *trans*-acting is located far from the target gene. Both variants have different influence on the levels of expression. Individuals with the G variant of the *cis* relationship have a higher expression and the same with individuals with the T variant in the *trans* relationship. (Copied from Vivian G. Cheung and Richard S. Spielman doi:10.1038/nrg2630)

2.3. Genomics – Epigenomics

DNA methylation regulates gene expression and genetic variants are associated with gene expression too, therefore it is plausible that genetic variants may be related with DNA methylation levels. Less studied than the others pairwise combinations is the study of methylation Quantitative Trait Loci (methQTL) where the genetic variants are associated with the methylation levels. The studies performed in the last years (Bell et al. 2010; Banovich et al. 2014; Heyn et al. 2014) have demonstrated that a genetic-epigenetic association exists pointing to new molecular mechanism behind complex diseases. As in the eQTL analyses, they can be classified as *cis*-acting (1MB distance between the CpG site and the SNP involved in the relation) and *trans*-acting, although the last ones have still not be very extensive studied.

To analyze these relationships, the models are constructed using only two different scales at a time, for instances, gene expression or SNPs that have either continuous values for the level of expression or categorical variables in the case of the SNPs indicating overexpressed or underexpressed genes depending on the allele. Same idea when the input variables are the CpGs with the difference that CpGs can also be measured in a continuous scale indicating an overexpressed or underexpressed genes with higher or lower levels of methylation. Normally, the aim of these analyses is to determine genes using SNPs or CpGs that may act as risk factors, mediators, confounders or effect modifiers. But at present, studies considering CpGs as the entities in a continuous scale and SNPs in a categorical scale indicating higher or lower levels of methylation depending on the SNP allele are also established. Different statistical methods to implement these pairwise combinations can be used, including linear regression or correlation.

2.4. Statistical methods for pairwise analysis

To assess correlations between two continuous variables such as in the study of epigenomics – transcriptomics pairwise, the typical method used is Pearson correlation coefficient. It was developed by Karl Pearson from a related idea introduced by Francis Galton in 1880s (Stigler 1989). This measure checks the linear correlation between two continuous variables.

Definition of Pearson correlation:

$$\rho_{X,Y} = \frac{cov\ (X, Y)}{\sigma_X, \sigma_Y}$$

where, cov is the covariance and σ_X, σ_Y are the standard deviation of *X* and *Y* respectively.

The correlation coefficient can take values from -1 to 1. A value of 1 implies a perfect positive correlation between *X* and *Y* while -1 implies a perfect negative correlation. A value of 0 means that there is no correlation. This coefficient belongs to a parametric test, requiring that the distribution of the variables follows a normal distribution. When it is not possible to assess the normality assumption, for example in the case of DNA methylation, Spearman's rank correlation coefficient (non-parametric test) can be used. It was developed by Charles Spearman (Spearman 1904) and measure the statistical dependence between two variables.

Definition of Spearman correlation:

Considering a sample of size *n* and being the *n* raw scores x_i, y_i

$$\rho_{X,Y} = 1 - \frac{6\sum d_i^2}{N(N^2 - 1)}$$

where $d_i = x_i - y_i$ is the differences between ranks. The values of the coefficient are interpreted as in the Pearson correlation.

For the analysis of eQTLS and methQTLs, the most popular method to identify them is through linear regression models. The linear regression modeling is an approach to assess the relationship between a dependent continuous variable *Y* (response variable) and one or more independent variables denoted as *X* (predictors). When only one predictor variable is used, we name it as simple linear regression model and multiple when more than one is used.

Definition of linear regression model:

Consider a data set where $y = (y_1, \dots y_n)^t$ is the response variable and $x = (x_{1j}, \dots x_{nj})^t\ j = 1, \dots p$ are the predictors, the model takes the form:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i,\ \ i = 1 \dots n$$

where *n* is the sample size, *p* the number of predictors and ε_i the error variable (residuals).

To estimate the parameters, it is normally used the ordinary least square (OLS) estimator that minimizes the sum of squared errors with the assumption that the sum of errors ($\sum \varepsilon_i$) is equal to 0. An example of how to obtain the regression model for a simple linear regression is shown:

$$min\left(\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n}(y_i - (\alpha + \beta_1 x_{i1}))^2\right)$$

To estimate *α* and *β*, we obtain the solution of the derivative conditioned to each parameter:

$$\frac{\partial}{\partial\alpha}\left(\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n}(y_i - (\alpha + \beta_1 x_{i1}))^2\right) = -2\sum_{i=1}^{n}(y_i - (\alpha + \beta_1 x_{i1}))^2 \qquad (1)$$

$$\frac{\partial}{\partial\beta_1}\left(\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n}(y_i - (\alpha + \beta_1 x_{i1}))^2\right) = -2\sum_{i=1}^{n}(y_i - (\alpha + \beta_1 x_{i1}))^2 \qquad (2)$$

To obtain the minimizing point, (1) and (2) are derivate and set to 0:

$$\sum_{i=1}^{n} y_i - n\alpha - \beta_1 \sum_{i=1}^{n} x_i = 0 \qquad (3)$$

$$\sum_{i=1}^{n} x_i y_i - \alpha \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \qquad (4)$$

Solving (3) and (4), the parameters are obtained as:

$$\alpha = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i\right) = \bar{y} - \beta_1 \bar{x}$$

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

where $\bar{y}$ and $\bar{x}$ are the mean of *x* and *y*, S_{xy} is the covariance between *x* and *y*, and S_{xx} is the variance of *x*.

To apply linear regression models, the following assumptions need to be verified: ***linearity*** (the mean of the response variables is a linear combination of the parameters (regression coefficients) and the predictor variables), ***homoscedasticity*** (same variance in the errors of the response variables)***, independence of errors*** (the errors are uncorrelated), and ***lack of multicollinearity*** in the predictors (the predictors cannot be correlated between them).

It is important to mention than when the response variable is binary or time-dependent, special cases from linear regression models are used: logistic regression and cox regression, respectively. Linear regression models are normally applied when the independent variables are the SNPs (categorical) and the response variable is continuous such as the gene expression levels (eQTLs) or the DNA methylation levels (methQTLs). Usually, one SNP at a time is analyzed to assume the lack of multicollinearity.

Chapter 3. Integrative –omics analysis

In the previous chapter, pairwise combination considering three –omics data was covered. These may provide some of the pieces of the puzzle of complex diseases showing new mechanism of the whole human genome, but the processes that happen in our body are much complex. Going one step further in integrative analysis to consider all the –omics data type together is a must. An overview of the statistical methods applied in –omics data integration analysis and its main challenges are described is this chapter.

3.1. Introduction to data integration

The concept of data integration can have numerous meanings: Lu et al. (2005)defined data integration in the context of functional genomics as the process of statistically combining diverse sources of information from functional genomics experiments to make large-scale predictions. Hamid et al. (2009) explained the data integration in a much broader context where it includes the fusion with biological domain knowledge using a variety of bioinformatics and computational tools and lastly Kristensen et al. (2014) and Ritchie et al. (2015) introduced the concept of data integration as a system biology approach. Kristensen et al. remarks that the principles of integrative genomics are based on the study of molecular events at different levels on the attempt to integrate their effects in a functional or causal framework. Ritchie et al. remarks that the complete biological model is only likely to be discovered if the different levels of genetic, genomic and proteomic regulation are considered in an analysis.

Based on these ideas, statistical methods are emerging specifically for –omics integrative analysis. Some examples in the literature have lastly explored the combination and integration of –omics data. Gibbs et al. (2010) combined both eQTLs and methQTLs in human brain. Bell et al. (2011) combined DNA methylation patterns with genetic and gene expression in HapMap cell lines or Wagner et al. (2014) combined also three data sets, DNA methylation, genetic, and expression in untransformed human fibroblasts. However, any of these analyses have combined more than 2 –omics data in the same model at the same time, mainly because of a lack of methodology to deal with the challenges arising with the implementation of integrative –omics analysis.

3.2. Challenges of -omics integrative analysis

While integrative –omics analysis will allow us to explore new questions and discover new findings, numerous challenges arise such as heterogeneity of the data, huge dimensionality, n << p problem, high correlation. At the individual data sets level, an exhaustive quality control, descriptive studies, and an estimation of missing values need to be conducted with a special attention since the whole analysis will depend on how good is done this process. When dealing with the ***huge heterogeneity*** between –omics data sets (SNPs are measured as categorical variables where 0, 1, and 2 are representing the number of variants while CpGs are measured in a continuous scale which is different from the gene expression continuous scale) numerous difficulties are attached. Thus, to be able to model them appropriately it is crucial to know in detail the scale structure of the data together with the biological meaning of each –omics and their relationships. Another main challenge is due to the ***high dimensionality*** of the data as millions of data per each data set are determined in the same individuals. Therefore, the necessity of performing data reduction in order to obtain the most relevant results appears. But even with data reduction, the huge amount of independent variables will be always smaller than the number of individuals ***(n << p)*** and it is also a problem to deal with. Consequently from the dimensionality and (n << p), statistical power becomes an issue and correction by multiple testing increase dramatically. To fix these issues, filtering is performing before analysis that may facilitate the integration in a smaller subset. This filtering can be done in a biological way, such as the one carry out by Biofilter (Bush et al. 2009) that uses public information from GWAS; or in a statistical way, through different statistical methods such as Principal Component Analysis (PCA) or Factor Analysis (FA) which are explained later. Another possibility of filtering is for example reducing the number of SNPs by Linkage Disequilibrium (LD) or CpGs that belong to the same CpG islands. These last approaches are also used to avoid ***high correlated*** data which is also a challenge to deal with. The problem with filtering is that it can exclude functional markers and lose important information. Nevertheless, the majority of statistical methods cannot be applied because of multi-collinearity (high correlation). So, it will be necessary or filtering before analysis assuming that some information may be lost or finding the proper statistical method to select the most relevant information.

Apart of the methodological challenges presented before, ***interpretation***, ***replication*** and ***validation*** of this complexity are also important challenges. After integrating the –omics data, normally a huge amount of results are generated and ways for interpretation are needed. Also, a way of controlling the possible identification of false positives association behind is needed. At the end, the results have to be trustable and understandable and the replication becomes an issue. If the findings derived from a single –omics analysis are difficult to replicate in an independent data set, the idea of replicating the combination of more than two –omics becomes almost impossible. So, new ideas for replicating and validating are needed.

3.3. Statistical methods for –omics integrative analysis

To perform integrative analysis, different strategies can be applied, one aims to divide data analysis into multiple steps, and signals are enriched with each step of the analysis. Another is to combine more than two –omics data sets simultaneously in the same model. Consequently, multivariable approaches need to be taken into account to face the challenges mentioned before.

To reduce data dimensionality, PCA is a method that converts a set of observations into a set of values of linearly uncorrelated variables called principal components (PCs). The new values generated retain most of the observable information based on the correlation between the original variables.

Definition of PCs

Considering a sample of *n* observations on a vector of *p* variables $x = \left(x_1, x_2, \dots, x_p\right)$ the first PC of the sample is defined by the linear transformation:

$$z_1 = \boldsymbol{a_1}^{\boldsymbol{T}}\boldsymbol{x} = \sum_{i=1}^{p} a_{i1}x_i$$

where the vector $\boldsymbol{a_1} = \left(a_{11}, a_{21,} \dots, a_{p1}\right)$ is chosen such that $var[z_1]$ is maximum.

Likewise, the *k*th PC is defined equally subject to

$$cov[z_k, z_l] = 0\, for\; k > l \geq 1$$

$$\boldsymbol{a_k}^T\boldsymbol{a_k} = 1$$

FA is related to PCA in that it is used to describe the variability among observed, correlated variables. This variability is recovered in what is called factor where multiple observed variables have similar patterns of responses because they are all associated with a latent variable.

Definition of FA

Considering the same vector of p variables $x = \left(x_1, x_2, \dots, x_p\right)$ as in the PCA, they can be expressed as linear functions and an error term, that is

$x_1 = \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1m} f_m + \varepsilon_1$k

$$x_2 = \lambda_{21} f_1 + \lambda_{22} f_2 + \cdots + \lambda_{2m} f_m + \varepsilon_2$$

$$\vdots$$

$$x_p = \lambda_{p1} f_1 + \lambda_{p2} f_2 + \cdots + \lambda_{pm} f_m + \varepsilon_p$$

Where λ_{jk} are constants called factor loadings, ε_j are the errors and $f_1, f_2, \dots, f_m$ are the factors. FA attempts to achieve a reduction from p to m, while the number of PCs are the same as the p variables. Both FA and PCA are used in single data sets to reduce dimensionality.

Also related with PCA, canonical correlation analysis (CCA) is important. In this case the application is in a two vector of variables *X* and *Y*, and it investigates the overall correlation finding linear combinations of the two sets of variables.

Definition of CCA

Considering $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_m)$ two vectors of random variables, it is defined the cross-covariance $\sum_{\boldsymbol{XY}} = cov\ (x, y)$ as a matrix where $cov\ (x_i, y_i)$ is the covariance for (x_i, y_i). CCA look for vectors a and b that maximize

$$\rho = \frac{a' \sum_{XY} b'}{\sqrt{a' \sum_{XX} a} \sqrt{b' \sum_{YY} b}}$$

The optimal linear combination of variables from the sets *X* and *Y* are called canonical vectors and are given by:

$$a = \Sigma_{xx}^{-1/2} u \qquad b = \Sigma_{yy}^{-1/2} v$$

From which the new variables $\eta = a'x$ and $\theta = b'y$ are obtained. They are called canonical variables or latent variables.

All these methods are more exploratory than hypothesis-testing, the PCA and FA work in linear combinations with one set of variables at a time and the CCA works in the linear combination of two sets of data. Moreover all of them generate thousands of variables without selection. They are useful in the context of filtering but not in the performance of –omics integrative modeling.

For more details look at Jollife (2002) book. Therefore, some statistical applications have been proposed for –omics integrative analysis.

Multiple factor analysis (MFA) is an extension of PCA where more than one variable can be studied. The idea is to find whether a group of individuals is described by a set of variables structured in groups. It was proposed by Brigitte Escofier and Jérôme Pagès in 1980s (Escofier & Pagès 1994). It is based on the computation of a PCA in each data set and then look for common factors. An example of an application in the contexts of –omics integration is shown in (de Tayrac et al. 2009). They focus on a study combining the genome and the transcriptome of gliomas. This method is a way of integrating data, but do not provide specific relationships between each data set. Another extension from the approach showed before is the sparse canonical correlation analysis (SCCA) that is an extension of CCA. SCCA allows the analysis of two sets of variables in order to establish the relationship between them. The idea behind is to add parameters λ_u and λ_v for variable selection to the vectors *u* and *v*. The entire algorithm is explained in detail in (Parkhomenko et al. 2009) with an example of application in genomic data integration. Another example of new implementations is the multivariate partial least squares (PLS) regression (Wold et al. 1984) that is a statistical method that support a relation to PC regression which is based on PCA. PLS is used to find relationships between two matrices (X and Y). The algorithm proceeds iteratively, extracting the linear combination of predictor variables that best describe the response variables. An example applied to microarray data is shown in (Palermo et al. 2009). These methods are based on frequentist statistics techniques, but there are others based on machine learning approach or Bayesian statistics that are not introduced is this thesis.

Even though, these methods have good properties, one goal of integrative analysis adopted in this thesis is to determine entities (i.e. genes) using at least two –omics integrated in the same model. For that, penalized regression methods have very good properties that overpass all the challenges aforementioned.

Chapter 4. From standard regression to penalized regression methods

In Chapter 2, I comment about the application of linear regression models to relate a variable response *Y* with *p* variables predictors X_1, X_2 ,... , X_p . In this model, the estimates of the coefficients are based in minimizing the sum of the squared error producing unbiased estimators. The bias of an estimator is defined as the difference of the expected value and the true value of the parameter. When unbiased, this difference is zero. Therefore, the mean squared error (MSE) measure how well the estimate is. It is defined as:

$$MSE = \frac{\sum_i (y_i - f_i)^2}{n - p - 1}$$

Where f_i is the estimated model of y_i.

For a given solution x_0,

$$MSE = E[(Y - f(x_0)]^2 = E\varepsilon^2 - E\big[[f(x_0)]^2 + \big[f(x_0) - E\big[[f(x_0) - E[f(x_0)]]^2\big]$$

$$= \ noise + \ bias^2 + variance$$

Where ε is the residual error from the adjusted model.

So, among unbiased estimators, minimizing the MSE is equivalent to minimizing the variance. Consequently, penalized regression methods sacrifice a little bias to reduce the variance of the predicted values through a shrinkage factor improving predictions overall as is represented in Figure 1.4.1.

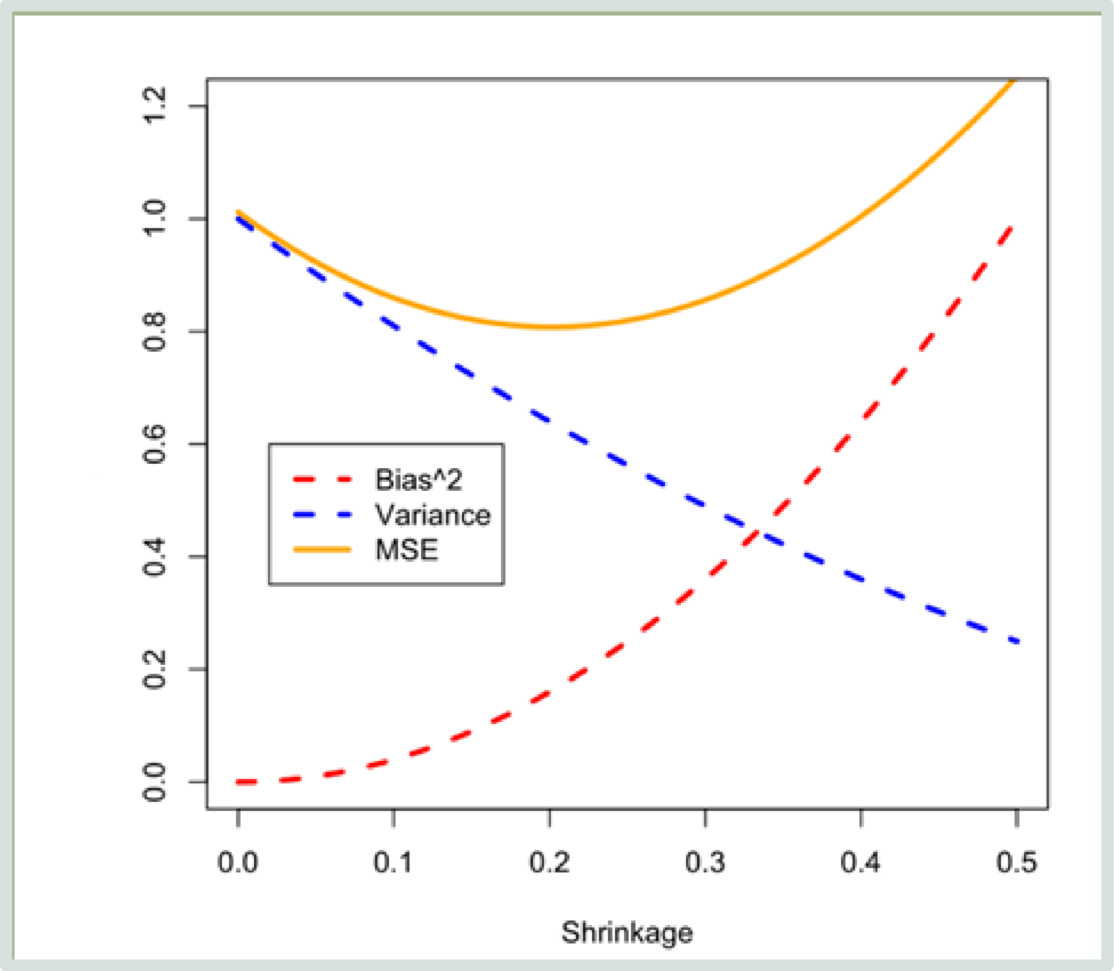


Figure 1.4.1. Graphical representation of the relationship between the shrinkage factor and the bias, variance and MSE. The bias estimator is increasing while the variance is decreasing when the amount of shrinkage increase. The optimal result have to maintain the minimum MSE as possible. (Copied from Stanford university webpage, Jonathan Taylor presentation on penalized models)

Penalized regression methods have very good properties for high throughput –omics integrative analysis. They can deal with the majority of the challenges listed in Chapter 3: they can be applied when the number of parameters is much higher than the number of samples, they produce sparse models to be interpretable, they allow for the use of different scale variables in the same model, so more than two –omics can be analyzed at the same time in the same model, and one of the most important properties in analyzing high-throughput data, they can deal with highly correlated variables.

The principle penalty functions that have been proposed are the l_1 norm solved by the Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Tibshirani (1996), the l_2 norm solved by ridge regression proposed by Hoerl & Kennard (1970), and the combination of the l_1 and l_2 norm solve by Elastic Net (ENET) proposed by Zou (2005).

Definition of the methods

Considering a multiple linear regression model:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \;\; i = 1 \ldots n$$

The estimators for LASSO, ridge and ENET are defined as:

$$\hat{\beta}_{lasso} = argmin \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - x_i\beta)^2 + \lambda_{lasso} \sum_{j=1}^{p} |\beta_j| \right\}$$

$$\hat{\beta}_{ridge} = argmin \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - x_i\beta)^2 + \lambda_{ridge} \sum_{j=1}^{p} \beta_j^2 \right\}$$

$$\hat{\beta}_{enet} = argmin \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - x_i\beta)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\}$$

The amount of shrinkage is determined by the parameters: λ_{lasso}, λ_{ridge}, λ_1 and λ_2 In the case of the LASSO and ENET, the values will cause shrinkage of the estimates of the regression towards 0. In the case of the ridge, the estimates of the regression will never be 0. This is the reason why in this thesis LASSO and ENET are the only penalized regression methods applied since, we are interesting in variable selection methods to obtain sparse results. Regarding the behavior of these three penalty functions, a graphical representation is shown in Figure 1.4.2 in a two-parameter case *β₁* and *β₂.* The shapes in the figure belongs to the constraints: $\sum_{j=1}^{p} |\beta_j| \leq t$ for LASSO, $\sum_{j=1}^{p} \beta_j^2 \leq t$ for ridge and the combination of both for ENET. As a consequence of the shapes LASSO and ENET are likely to perform variables selection *β_1=0* and/or *β_2=0.*

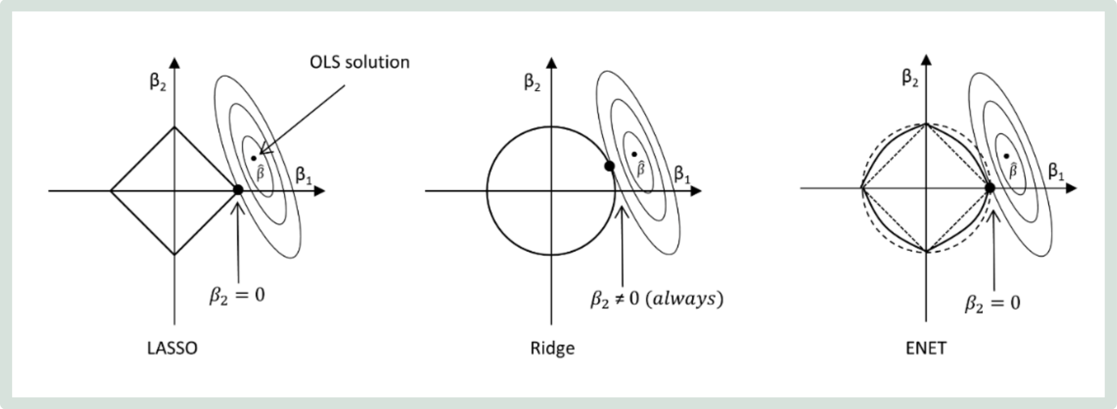


Figure 1.4.2. Graphic representation of the different penalty functions. The rhombus, circle and oval shapes represent the LASSO, Ridge and ENET constraint, respectively. The eclipses represents the penalized likelihood contours from the OLS solution ($\hat{\beta}$) and the dots are the penalized likelihood solution tangent to the constraints. If the likelihood contour first touch the constraint at point zero, the estimate is zero and the variable is not selected. In case of ridge, the eclipses can never touch the point zero due to the circle shape.

To select the optimal penalization parameter, k-fold cross validation (CV) (Trevor Hastie, Rob Tibshirani and Jerome Friedman 2001) is used. The measure used in the CV normally is the MSE, but others can be used such as deviance, area under the curve, etc…In Figure 1.4.3 and Figure 1.4.4 an example is shown when applying LASSO to a multiple linear regression model for explaining gene expression levels by several SNPs. Figure 1.4.3 shows the selection of λ_{lasso} by CV. Each red dot represents the value of the MSE per each value of λ_{lasso}. The optimal value is when the MSE is minimum and correspond to a $\lambda_{lasso} = 0.89$ selecting 5 variables different from 0.

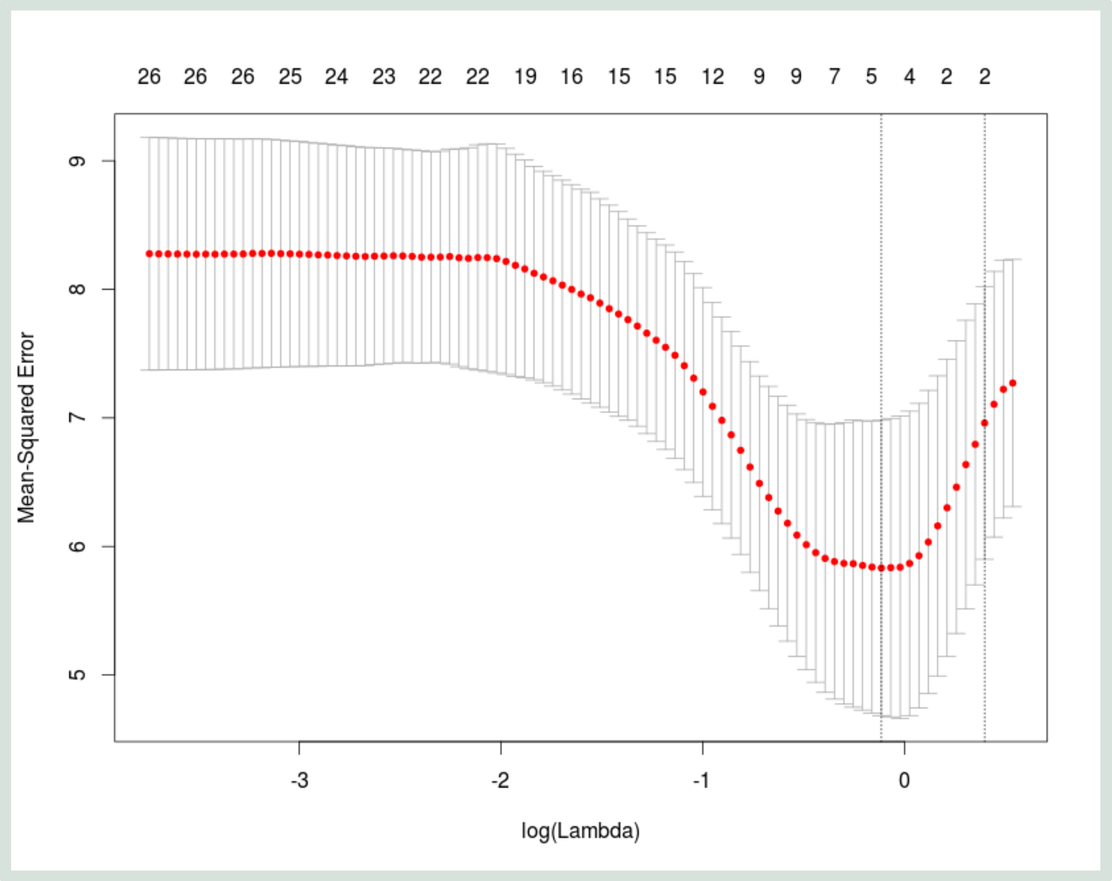


Figure 1.4.3 Values of MSE with the confidence interval for the different values of lambda ($\boldsymbol{\lambda_{lasso}}$). The y-axis represents the values of MSE, the down x-axis represents the values of lambda (λ_{lasso}) and the up x-axis represents the number of variables different from 0.

Figure 1.4.4 shows the shrunken values of the coefficients of the regression model when λ_{lasso} vary and it is observable that when $\lambda_{lasso} = 0.89$, only 5 color lines are different from 0, that are the values selected from the LASSO with the optimal lambda.

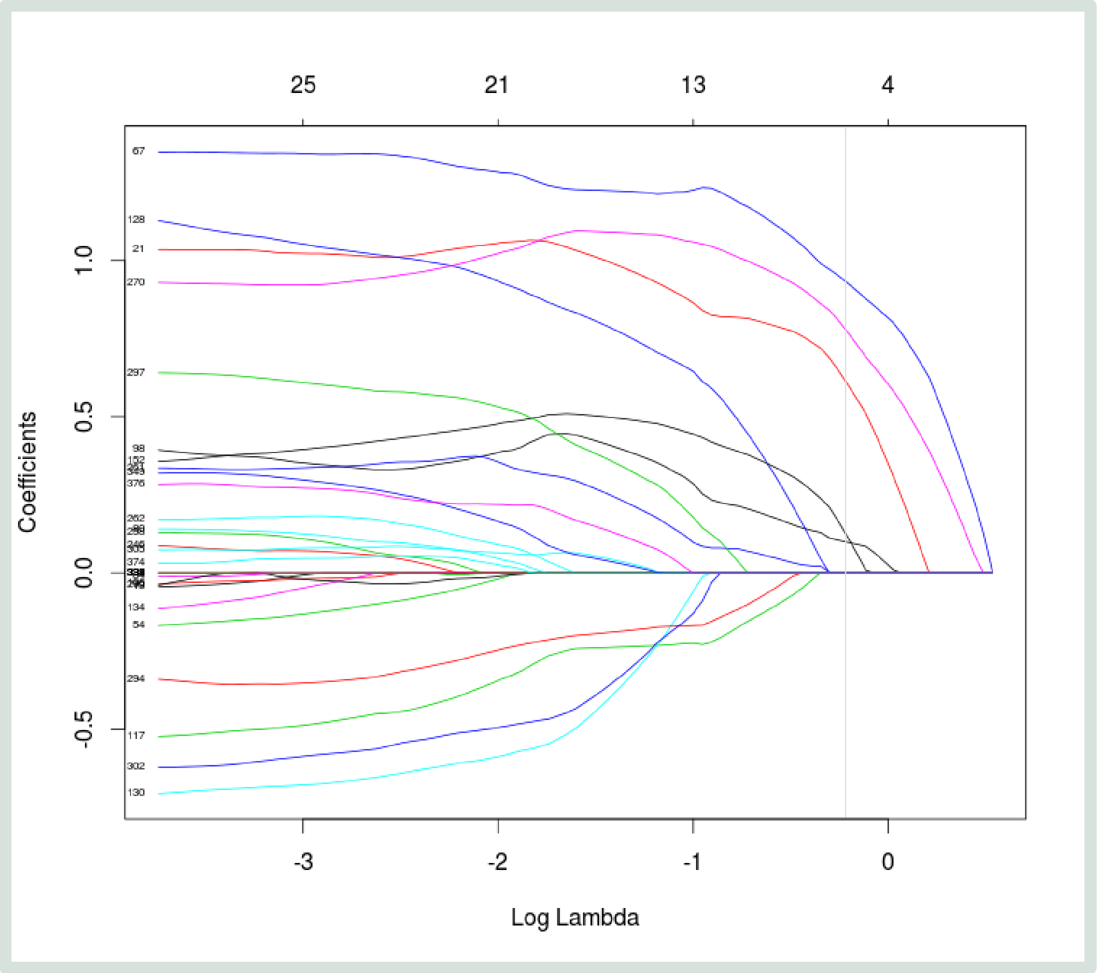


Figure 1.4.4 Representation of the shrunken coefficients of the regression el for the different values of lambda. The y-axis represents the value of the coefficients, the down x-axis represents the values of lambda (λ_{lasso}) and the up x-axis represents the number of variables different from 0.

Penalized regression methods have been applied in GWAS studies context (Wu et al. 2009; Ayers & Cordell 2010; Van Eijk et al. 2012; Chen et al. 2010; Zhou et al. 2010) where penalized logistic regression was applied. Also, a recent evaluation of the LASSO and ENET in GWAS studies has been published (Waldmann et al. 2013), and we also have previously used penalized regression methods in a candidate pathway analysis where we assess genetic variation in the *TP53* pathway and Urothelial Bladder Cancer (UBC) risk (Pineda et al. 2014) [Appendix 1]. This work was developed as part of my Master thesis performed during the first year of my PhD fellowship. Briefly, we investigated a total number of 184 tagSNPs in a case/control study where we applied first a classical statistical analysis using logistic regression to assess individual SNPs association and second the LASSO penalized logistic regression analysis to assess all the SNPs simultaneously. Finally, penalized regression methods have been applied in integrative analysis (Mankoo et al. 2011), where penalized Cox regression was used.

Chapter 5. Introduction to bladder cancer and data types

5.1. Cancer Epidemiology

Cancer is the leading cause of death worldwide (Ferlay et al. 2013). There were 14.1 million new cancer cases, 8.2 million cancer deaths, and 32.6 million people living with cancer in the last estimation from Globocan (http://globocan.iarc.fr/) for the period of 2012. Notable are the differences observed between sex: the overall age standardize rate (ASR) cancer incidence is around 25% higher in men than in women (205 vs. 165 per 100,000); and also among regions: from Western Africa (95.3 per 100,000) with the lowest incidence rate to Australia (318.5 per 100,000) with the highest rate (Figure 1.5.1).

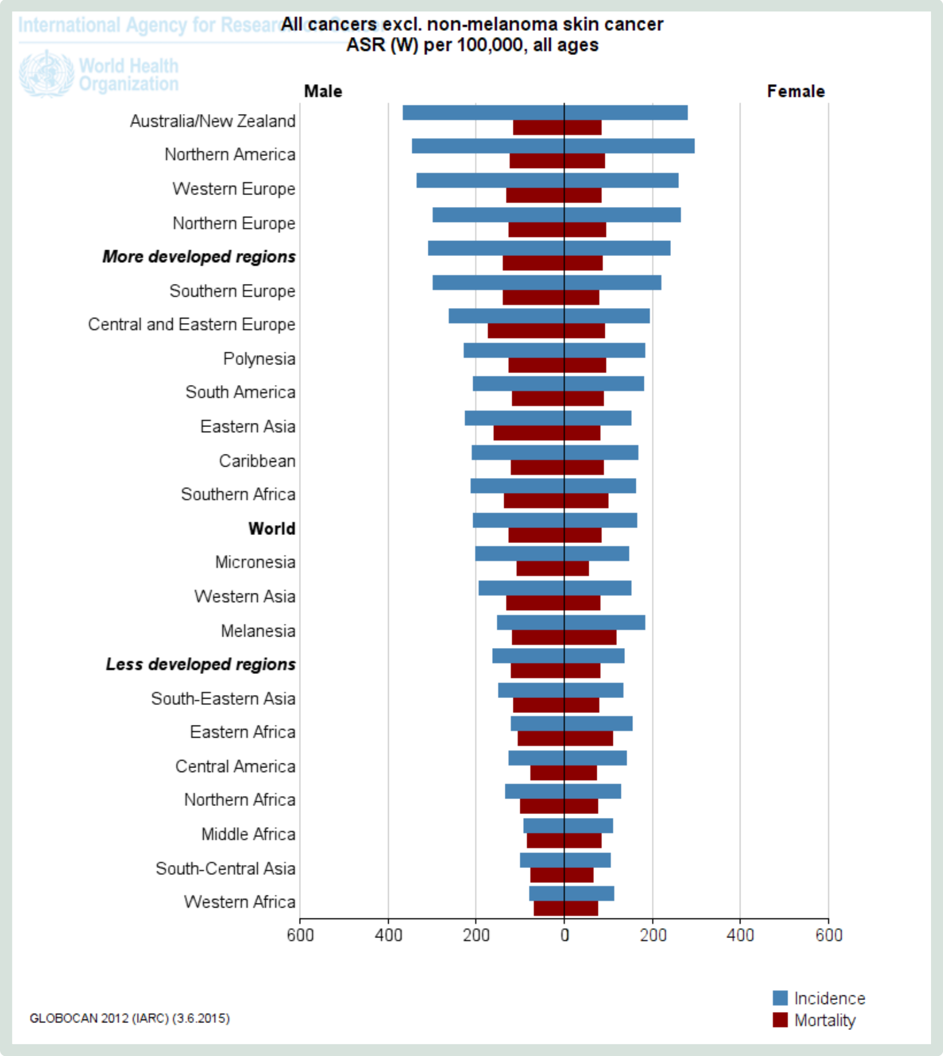


Figure 1.5.1. Incidence and mortality rates for all cancers separated between males and females in different regions worldwide. (Extracted from Globocan 2012)

5.2. Bladder cancer epidemiology

The present work uses a –omics dataset from bladder cancer patients. This represents one of the major types of cancer with 429,793 new cases and 165,084 deaths according to the estimation from Globocan (2012). The ASR varies across regions, with a higher rate in Europe with approximately 12 per 100,000. The highest incidence rate in Europe is shown by Belgium, Spain being in the 5th position. In terms of gender, bladder cancer also affects more men than women (9 vs. 2.2 per 100,000 the ASR respectiAffectvely) (Figure 1.5.2).

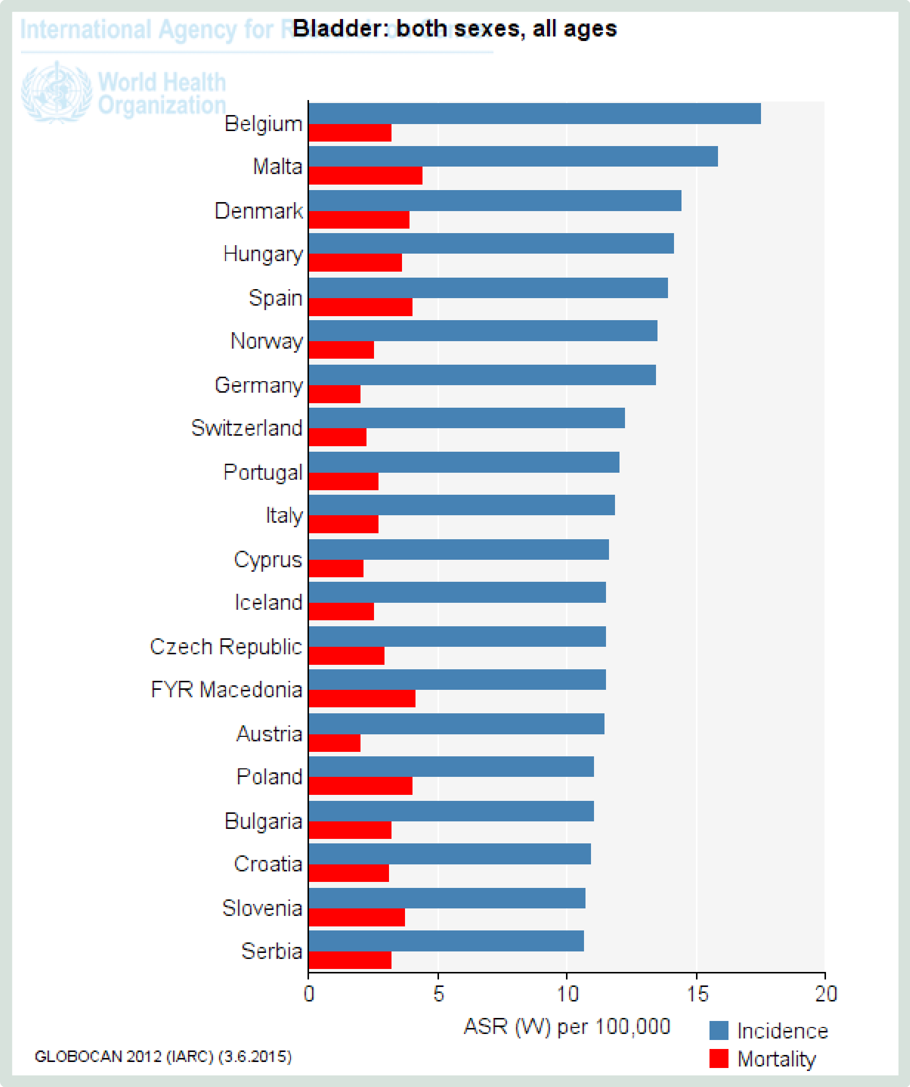


Figure 1.5.2. Incidences and Mortality ASR per 100,000 for the 20th highest in Europe for both sexes. (Extracted from Globocan 2012)

Bladder cancer is an important public health problem in Spain, mainly among men being the 5th most frequent cancer (ASR= 13.9 per 100,000) but with a huge difference between the incidences rates for men (ASR= 26.0) and women (ASR = 3.7) with a gender man:woman ratio of 7:1, in contrast with the ratio 3:1 in the westernized world.

5.3. Bladder cancer tumorigenesis

Bladder cancer encompasses various types of cancer according to their morphology, urothelial bladder cell carcinoma (UBC) being the most common occurring in up to 90% of all bladder cancer patients. UBC is further subtype in three groups according to their grade of differentiation (G) and stage (T): 1) low-risk, papillary, non-muscle invasive (NMI) tumors (60-65% of all UBC), 2) high-risk NMI (15-20% of all UCB), and 3) muscle invasive (MI) (20-30% of all UCB). Supporting these morphological subtypes, differential genetic pathways were identified. While deletion of both arms of the chromosome 9 is an initial step in bladder carcinogenesis as similarly frequent in both subtypes, somatic mutation in *FGFR3* are more frequent in low-risk NMI tumors, while mutations in *TP53* and *RB* pathways are mainly involved in high-risk NMI and MI (Wu 2009). Mutations in *PIK3CA* are a common event that can occur early in NMI supporting the hypothesis of different molecular pathways (López-Knowles et al. 2006).

5.4. Bladder cancer etiology

Bladder cancer is a complex disease that involves environmental exposures and genetic factors for its development. Cigarette smoking, occupational exposures, arsenic, Schistosoma haematobium infection, some medications, and genetic variation are the major risk factors associated with the disease as reviewed recently in (Malats & Real 2015). Tobacco consumption is the best established environmental risk factor and also occupational exposure to aromic amines, polycyclic aromatic hydrocarbons, and dyes have been associated with bladder cancer risk (Samanic et al. 2006; Samanic et al. 2008). For genetic factors, one study conducted in Scandinavian twins population-based estimated that 31% of the total variance of bladder cancer is explained by genetic factors while non-shared environmental factors would explain the 67% (Lichtenstein et al. 2000). Even though there is no high-penetrance allele/gene, low penetrance genetic variants have been found associated with bladder cancer risk. *NAT2* slow acetylation and *GSTM1* null genotypes increase UBC risk and in addition, the interaction between tobacco and *NAT2* is also well established (García-Closas et al. 2006). In addition, polymorphism in these genes (*MYC, TP63, PSCA, TERT-CLPTM1L, TACC3-FGFR3, CBX6, CCNE1*) have been identified associated with bladder cancer risk thorough GWAS (Nathaniel Rothman et al. 2010).

5.5. Bladder cancer data and –omics assessment

The data used in this thesis in Chapters 1 and 2 of Part 3 come from the pilot phase of the SBC/EPICURO. This is a multicenter hospital-based case-control study conducted in Spain between 1998 and 2001. The pilot phase was implemented recruiting individuals in 2 hospitals in Spain (Hospital del Mar,Barcelona, and Hospital General de Elche) during 1997-1998 and included total of 70 patients newly diagnosed of a histologically confirmed UBC with available fresh tumor tissue from which tumor DNA and RNA were successfully extracted and used. Table 1.5.1 displays the characteristics of the individuals included in the pilot study. The majority were males (93%) and current (50%) or former (36%) smokers. Based on the disease subtypes, 45% of individuals had low-grade- NMIBC, 22% had high-grade NMIBC, and 29% had MIBC.

Table 1.5.1. Characteristics of the studied patients

Characteristics	N (%)
Total	72
Gender	
Male	67 (93)
Female	5 (7)
Age	
Mean (SD)	65.6 (9.5)
Min-max	41-80
Region	
Barcelona	31 (43)
Elche	41 (57)
Smoking status	
Non-smoker	8 (11)
Current	36 (50)
Former	26 (36)
Unknown	2 (3)
Tumor-stage*	
Low-grade-NMI	32 (45%)
High-grade-NMI	16 (22%)
MI	21 (29%)
Unknown	3 (4%)

* Risk group was defined according to the grade (G) and stage (T) characteristics.

Genomics and epigenomics data were available for 46 individuals and transcriptomics data for 43. The overlapping between the three –omics data was 27. Genomic data was assessed with SNP genotyping with the IlluminaHap 1M array, epigenomics with bisulphite Infinium Human Methylation 27 Bead chip Kit detecting CpG sites and transcriptomics with the measurements of the levels of gene expression with the Affymetrix DNA microarray Human Gene 1.0 ST Array. We dedicate the next part of the thesis (*PART 2*) to describe in detail each –omics data set and the preprocessing and quality control analysis we applied.

For the replication purposes, UBC tumor and blood data from the TCGA consortium (https://tcga-data.nci.nih.gov/tcga/) was used. Already preprocessed data (level 3) was downloaded with TCGA-Assembler (Zhu et al. 2014). The data was profiled for 905,422 SNPs with the Genome wide 6.0 Affymetrix array for tumor tissue and blood samples, 20,502 gene expression probes with the RNASeqV2 platform for tumor tissue, and 350,271 CpGs with the HumanMethylation450K Illumina array for tumor tissue. 238 individuals with overlapping data from the 3–omics measured in tumor tissue and 181 with overlapping data also from genomic blood samples contributed to replicate results from Chapter 2 - Part 3 and in the discovery phase of Chapter 3 - Part 3.

Chapter 6. Hypothesis, Objectives and Thesis organization

6.1. Hypotheses

This is mainly a methodological development endeavor based on the needs voluminous agnostic/exploratory studies require. While there is no a specific scientific hypothesis behind the –omics exploration, my thesis pretend to support the concepts that 1) integrative –omics studies is a tool to find new mechanisms to better characterize the complex genetic architecture of complex diseases and 2) the amount of –omics data generated needs from the development of appropriate methodological approaches to analyze them and overcoming the abovementioned challenges this field imposes.

6.2. Objectives

The general objective of this work was to dissect and fix the methodological challenges of –omics data integration by combining different –omics data sets (genomics, epigenomics, and transcriptomics) under the umbrella of systems biology to identify relationships between and within the different types of molecular structures.

The specific objectives:

1. To perform the integration of three –omics data measured in tumor tissue in a multi-step process where all possible pairwise combinations are considering.
2. To perform the integration in a multi-dimensional approach where three –omics are analyzed in the same model at the same time.
3. To perform the integration of four –omics considering different levels of source material (tumor and blood samples) by adapting the previous developed tool to a 2 Stage Regression approach.

6.3. Thesis organization

The thesis is organized in 5 parts. *PART 1* already presented an introduction to the –omics data integrative field and the resources upon which this thesis has been conducted. *PART 2* describes in detail the pre-processing of the data and the quality control applied to each of the 3 –omics data used in this thesis. *PART 3*, structured as 3 scientific manuscripts, addresses the specific scientific and methodological objectives of this thesis. Within this part, *Chapter 1* proposes a framework analysis for the integration of three –omics data based on a multi-step process integrating all the possible pairwise combinations. *Chapter 2* proposes an integrative model to jointly analysis 3 –omics data using penalized regression methods. *Chapter 3* proposes an integrative eQTL –omics multi-material level analysis considering tumor tissue and blood samples. Finally, the last two parts are a general discussion (*PART 4*) and the conclusions of the thesis (*PART 5*).

PART 2.

PRE-PROCESSING AND QUALITY CONTROL

The -omics data measures are subject to different noises and errors and a number of critical steps are required to preprocess the raw data. This part of preprocessing following of the appropriate Quality Control (QC) probably is the main and most important part of the entire integrative analysis. The different approaches to implement the preprocessing and QC are data type-depending and will differ over the –omics data types and the high-throughput technologies used. This is the initial stage of all the data integration process that will be follow of a basic analysis to visualize graphically and statistically the different data types. The integration process, and therefore the statistical approach to perform the integration analysis will be based on what it is identified in this stage. Also, in this stage any problem or anomaly of the data can be detected.

In this part, the preprocessing, QC and basic analysis is described from the three types of –omics data from the SBC/EPICURO project that are used in this thesis. Chapter 1 describes the genomics from blood and tumor tissue and a comparison between both measures. Chapter 2 describes the epigenomics data (DNA methylation) from tumor tissue and chapter 3 describes the transcriptomics data (gene expression) from tumor tissue.

Chapter 1: Genomics from blood and tumor tissue

This chapter describes the preprocessing and QC of the genomic data from the SBC/EPICURO project used in this thesis as well as its basic analysis. A comparison between the genomic data measured in blood and tumor tissue are also assessed in this chapter.

1.1. SNP genotype data from blood samples

SNPs were genotyped in blood samples using two different platforms, the Illumina HapMap 1M array and the Illumina HumanHap Omni Express array. A total of 1,046,990 SNPs were genotyped in 39 individuals with the first platform and 703,525 SNPs were genotyped in 16 individuals with the second array. The data generated by both Illumina array platforms were visualized and analyzed with BeadStudio software separately. For the first platform, since the number of individuals was quite small, we decided to obtain the genotype calling using the cluster file obtained when the same array was applied to germline DNA from 2,424 subjects included in the main SBC/EPICURO study. This cluster file was imported to the BeadStudio project and the cluster analysis was processed for all the SNPs generating a SNP matrix with the genotypes per individual and the information of each SNP (dbSNP name, variant, position and chromosome). For the second platform as the array was different from that applied previously in the SBC/EPICURO study and the sample size was very small, we used the cluster file from Illumina. In both cases, from Beadstudio the genotypes (AA, Aa, aa) were obtained in forward strand for those samples having a call rate higher than 90% and introduced to R software to perform the pre-processing and QC. First, the genotypes were transformed in numerical categories being 0 (the common homozygous), 1 (the heterozygous) and 2 (the variant homozygous). Second, the number of missing and the Minor Allele Frequency (MAF) was calculated. The categories are represented in Table 2.1.1 In the first column are represented the SNPs obtained from the first platform and in the second column the SNPs that are common for both platforms (547,068). For both array, the annotation was done using the UCSC hg19, NCBI build 37 to make them comparable and homogenize their position in the genome.

1.2. SNP genotype data from tumor tissue samples

SNPs were genotyped using also the Illumina HumanHap 1M array in tumor samples. A total of 1,047,101 SNPs were genotyped in 46 individuals. As in the genotyping in blood the genotype calling was performed using the cluster file from the same array applied to germline DNA from 2,424 subjects included in the main SBC/EPICURO study. The same pre-processing and QC was applied and the SNPs by MAF and missingness categories are represented in the third column of Table 2.1.1. The annotation was also done using the UCSC hg19, NCBI build 37 to make the array comparable and homogenize its position in the genome.

Table 2.1.1: Summary of SNPs in blood and tumor

	SNP blood	SNP blood	SNP tumor
	Illumina Hap 1M	Illumina Hap 1M +	Illumina Hap 1M
		Illumina HumanHap Omni	
	(n=39)	(n=39+16)	(n=46)
Nº SNPs	1,046,990	547,068	1,047,101
maf			
= 0.0	151,075	47,860	150,548
(0.01 – 0.2]	399,767	221,748	420,716
(0.2 – 0.4]	344,976	189,286	327,762
> 0.4	151,172	88,174	148,075
Nº missing			
No missing	982,017	510,884	488,288
5% missing	44,545	28,551	400,918
20% missing	11,318	3,809	147,732
> 20% missing	9,110	3,824	10,163

The overlap between SNPs in blood and tumor was 543,244 for 29 individuals. For all the analysis, SNPs that have a MAF > 0.05, < 20% of missingness, a LD ≠ 1 and less that two individuals with the variant allele to avoid an increase number of false positives were considered.

Based on the idea that tumors acquire frequent somatic alterations, a concordance analysis was performed to see whether the differences are enough significant to consider these two measurements as different –omics data sets. To perform this analysis, kappa weighted measurement was applied to obtain the disagreement between two SNPs (tumor vs. blood). Each pair is represented in a weighted matrix where cells located on the diagonal represent complete agreement, while cells one off the diagonal are weighted 1, and cells two off the diagonal are weighted 2. Kappa takes values from 0 to 1, being 0 total disagreement and 1 total agreement. An example for one pair is shown in Box 1 where the disagreement between the SNP measures in blood and tumor was kappa = 0.35.

Box 1. Example of the application of weighted kappa in a SNP pair

Blood Tumor	0 (AA)	1 (Aa)	2 (aa)	Total
0 (AA)	17	3	0	20
1 (Aa)	1	1	1	3
2 (aa)	0	0	0	0
Total	18	4	1	23

$$kappa = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}m_{ij}}, where$$

$$k = \ number\ of\ codes\ and$$

$$w_{ij}, x_{ij}\ and\ \ m_{ij}\ are\ the\ weight, observed\ and\ expected\ values\ respectively.$$

The expected values are:

	0 (AA)	1 (Aa)	2 (aa)	Total
0 (AA)	15.65	3.48	0.87	20
1 (Aa)	2.35	0.52	0.13	3
2 (aa)	0	0	0	0
Total	18	4	1	23

The weighted matrix is:

	0 (AA)	1 (Aa)	2 (aa)
0 (AA)	0.0	1.0	2.0
1 (Aa)	1.0	0.0	1.0
2 (aa)	2.0	1.0	0.0

$$\boldsymbol{kappa} \ = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}m_{ij}}$$

$$= 1 - \frac{(17*0+3*1+0*2+1*1+1*0+1*1+0*2+0*1+0*0)}{(15.65*0+3.48*1+0.87*2+2.35*1+0.52*0+0.13*1+0*2+0*1+0*0)}$$

$$= 1 - \frac{5}{7.7} = \mathbf{0.35}$$

After applying this measure to the whole set of overlapping SNPs (543,244), we found that there were disagreement in all the chromosomes. This result is expected due to the somatic mutations produced in the tumor. Figure 2.1.1 represents the kappa coefficient by chromosome in a reverse Manhattan plot and Figure 2.1.2 represents the percentage of disagreement by chromosome considering the number of SNP pairs with kappa ≤ 0.8 divided by the total number of SNP pairs in the chromosome.

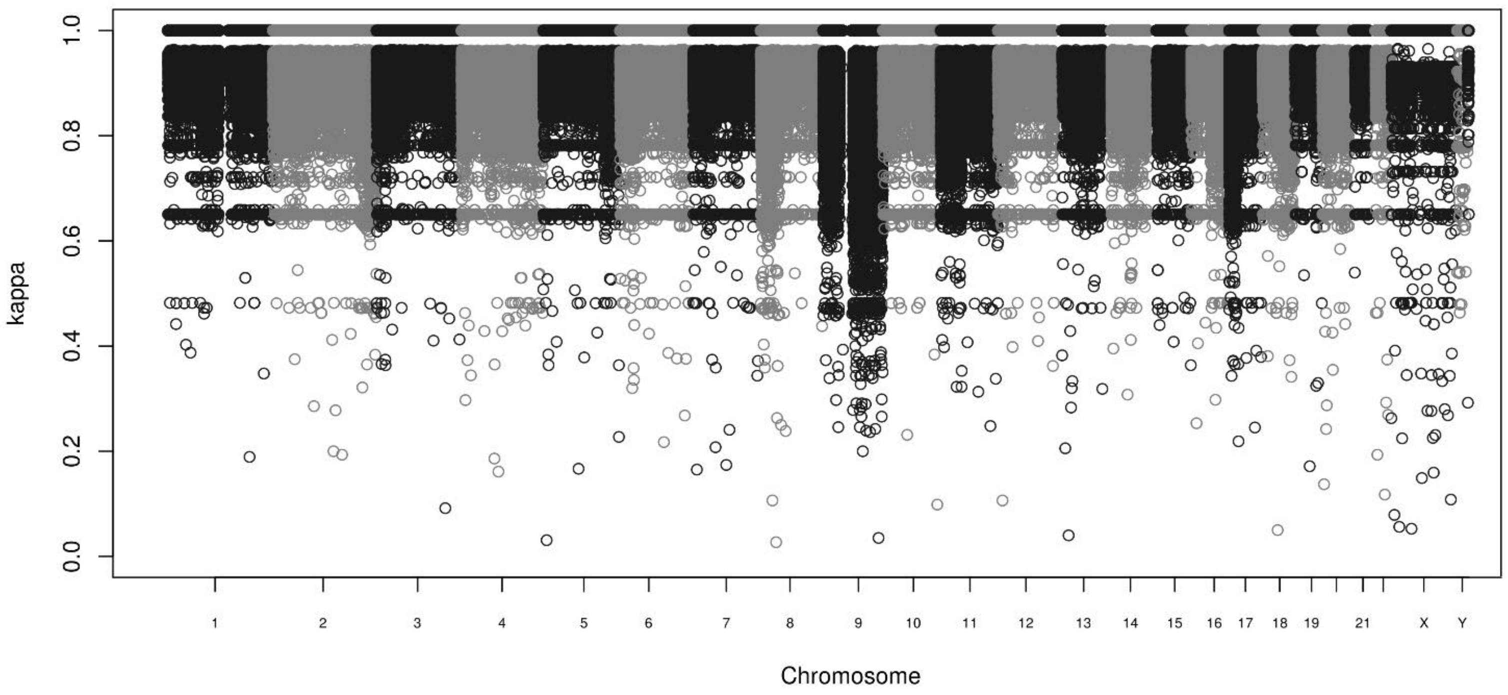


Figure 2.1.1 Kappa coefficient by chromosomes

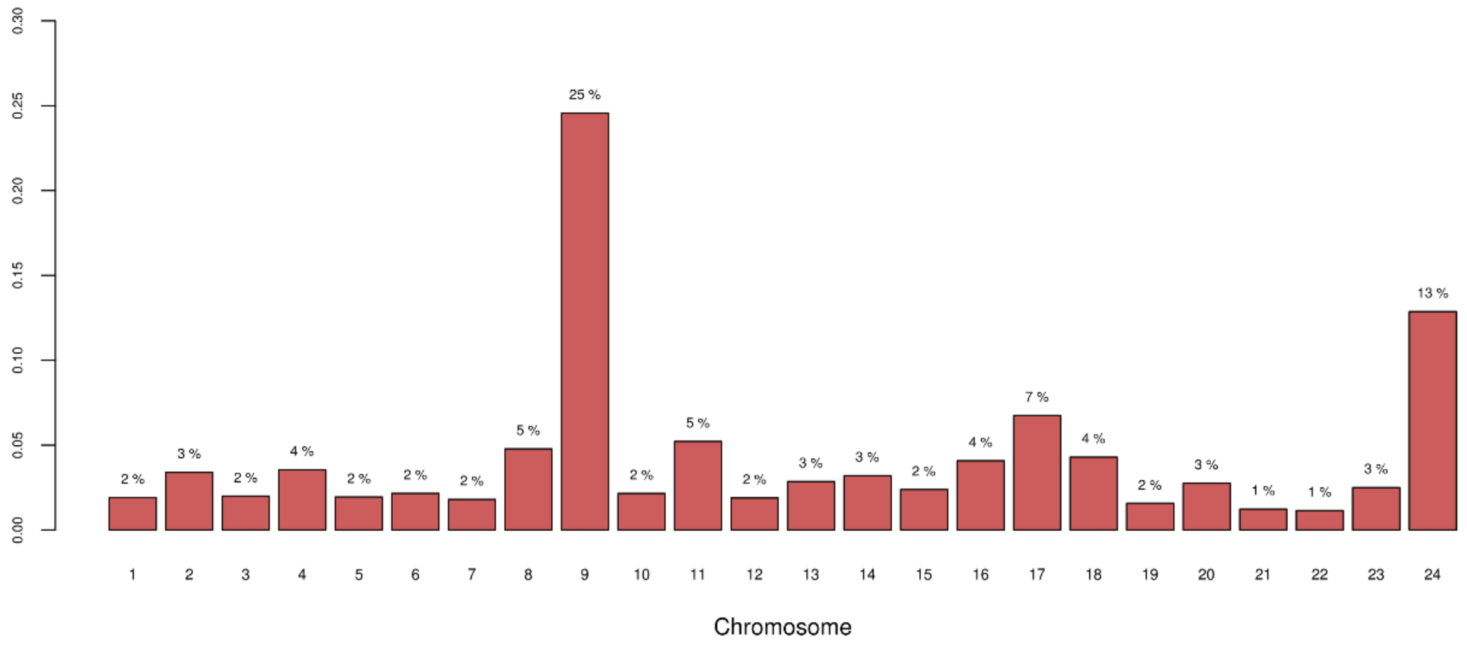


Figure 2.1.2. Percentage of disagreement by chromosome considering the number of pair SNPs with kappa ≤ 0.8 divided by the total pair SNPs in the chromosome.

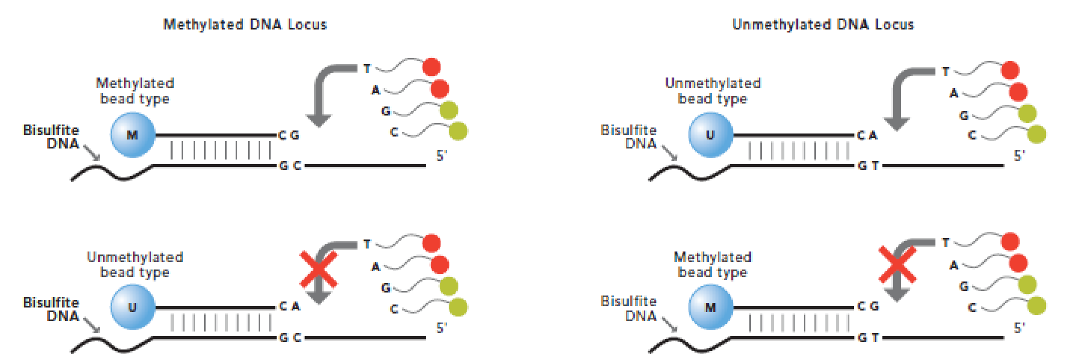
Chromosome 9 was the chromosome with the highest percentage of disagreement (25%) which can be explained with the early deletion of both arms of chromosome 9 in many UBC cases (Wu 2005). Chromosomes Y (13%), 17 (7%), 8 (5%) and 11 (5%) showed larger disagreement in comparison to the others. Deletions in the short arms of chromosomes 8 and 11 were associated with bladder tumor progression (Wu 2005). These results supported to consider the two measurements (tumoral genotypes and germline genotypes) as two different –omics data sets.

For chapters 1 and 2 in part 3 of this thesis, genomic measure in tumor was used. In chapter 1 univariable analyses was applied and a sample without missing was not required, but in chapter 2, multivariable models were applied requiring no missing values to avoid problems with a very small sample size. For this reason, we performed an imputation analysis using BEAGLE 3.3.2. with the method for inferring haplotype phase and sporadic missing data in unrelated individuals (Browning & Browning 2007).

Chapter 2: Epigenomics from tumor tissue

This chapter contains the preprocessing, QC, and exploratory single analysis of the epigenomic data from the EPICURO project.

DNA methylation was assessed in 46 tumor samples with the Infinium Human Methylation 27 BeadChip platform that quantitatively generate 27,578 CpG dinucleotides spanning 14,495 genes. To generate the CpGs, first an initial bisulfite conversion step is performed before the automated Infinium assay. Unmethylated cytosines are chemically deaminated to uracil in the presence of bisulfite, while methylated cytosines are refractory to the effects of bisulfite and remain cytosine. After bisulfite conversion, each sample is purified and applied to the BeadChips. To estimate the methylation status, two bead types are used that correspond to each CpG locus –one to the methylated (M) and the other to the unmethylated (U) state. Both bead types for the same CpG locus will incorporate the same type of labeled nucleotide, determined by the base preceding "C" in the CpG locus (Figure 2.2.1).



The Infinium Assay for Methylation detects methylation status at individual CpG loci by typing bisulfite-converted DNA. Methylation protects C from conversion (left), whereas unmethylated C is converted to T (right). A pair of bead-bound probes is used to detect the presence of T or C by hybridization followed by single-base extension with a labeled nucleotide.

Figure 2.2.1. Infinium assay for methylation. (Copied from Illumina: http://www.illumina.com/documents/products/appnotes/appnote_dna_methylation_analysis_infinium.pdf)

Then, the array is fluorescently stained and the intensities of the methylated and unmethylated bead type are measured with the β-values that are recorded for each locus in each sample via BeadStudio software. The β-value is defined as:

$$\beta = \frac{\max(M,0)}{\max(U,0)+\max(M,0)+100}$$

The maximum between signal intensity and 0 is used for *β* calculation to avoid the negative numbers caused by background subtractions, consequently, *β*-values rank between 0 (unmethylated) and 1 (methylated). The constant 100 was used to regularize the *β*-values when they were very small. *β*-value has a direct biological interpretation that corresponds to the percentage of methylated sites, but for analytical and statistical purpose, the *β*-value has severe heterocedasticity which impose a challenge in applying many statistical methods (Du et al. 2010); consequently *M*-value has been proposed as a logarithm transformation that is more statistically valid even though it does not have an intuitive biological meaning. The *M*-value is calculated as follows:

$$M = log_2\left(\frac{\max(M,0)+1}{\max(U,0)+1}\right)$$

M-value ranges between -∞ (unmethylated) and +∞ (methylated). In our study, *M*-values were used when applying linear regression models, while *β*-values were used in the rest of the analyses.

For the 46 tumor samples in EPICURO, we obtained the *β*-value from BeadStudio software with the detection p-values for the total number of 27,578 sites. CpGs with a detection p-value > 0.05 as Illumina recommended were rejected leaving 27,164 sites. Then, the CpGs with *β*-values < 0 or > 1 were also excluded yielding 26,634 sites. The Infinium HumanMethylation27 array detects some CpGs that are non-specific and map to genomic sequences and also some CpGs that cross-react with other sequences. (Chen et al. 2011) published a list of CpGs that are SNPs or cross reactive probes for this specific array that we used to filter CpGs in our analysis. A total number of 908 CpGs were SNPs and 2,985 were cross-reactive probes. Finally a total number of 23,034 CpGs were kept and used in chapter 1 and 2 in PART 3 of this thesis. The annotation was done using the UCSC hg19, NCBI build 37 to make the array comparable and homogenize its position in the genome.

An exploratory analysis was performed to inspect the patterns of DNA methylation according to *β*-value and *M*-value. In the Figure 2.2.2 the distribution of *β*-values and *M*-values is represented according to CpGs located in the autosomal chromosomes for both sexes and the X-chromosome in females. Table 2.2.1 and Table 2.2.2 shows the distribution of the *β*-values and *M*-values respectively according to three categories of methylation levels (low, medium and high).

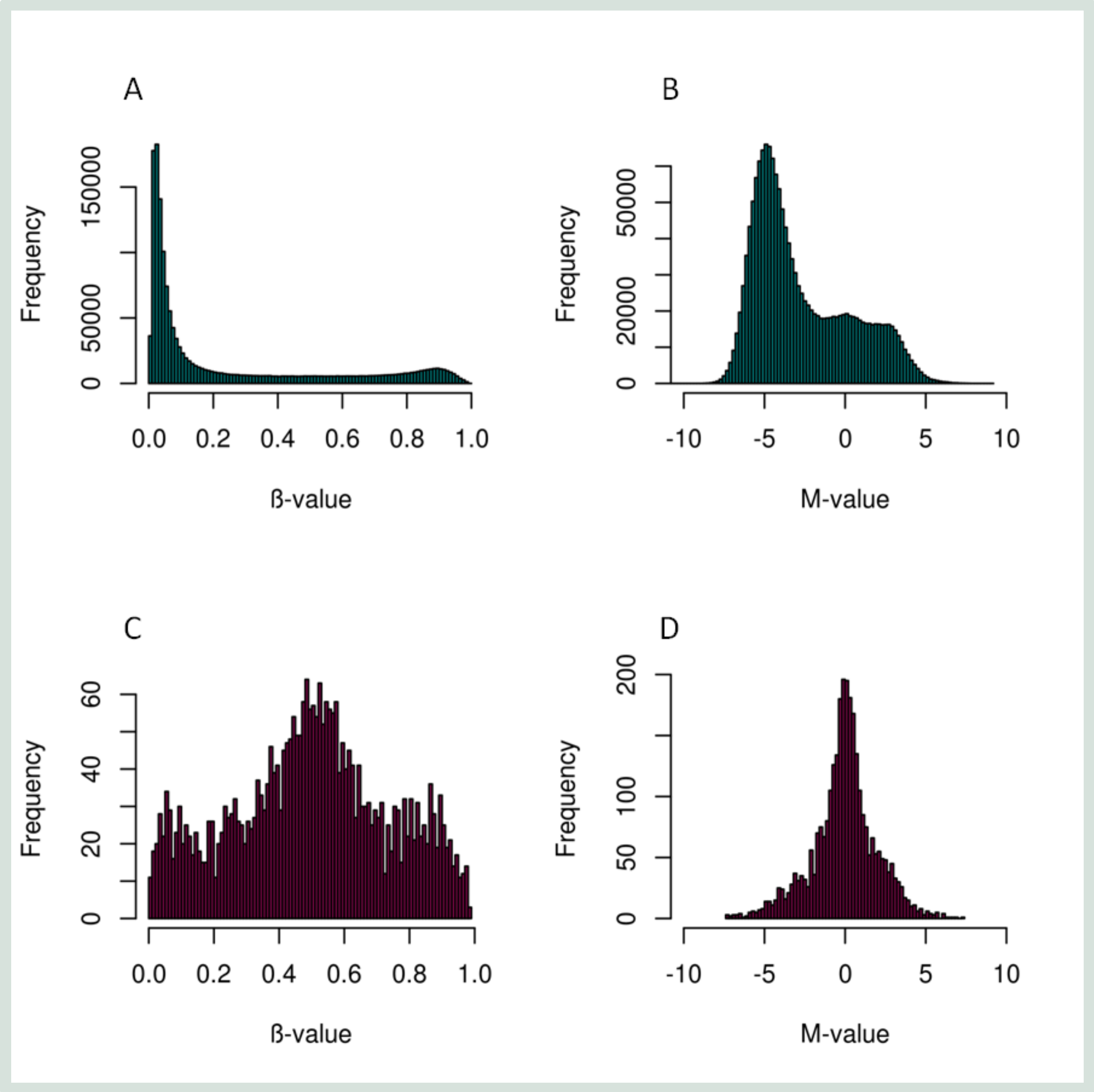


Figure 2.2.2. Distribution of the DNA methylation data. *β*-values for autosomal chromosomes (A) and X-chromosome (B) only females. *M*-values for autosomal chromosomes (C) and X-chromosome (D) only females.

The differences observed in our data for autosomal chromosome and X-chromosome in females are concordant to the already known patterns due to the X-chromosome inactivation. This is a mechanism that silences the majority of the genes on one X chromosome in each female cell (Carrel & Willard 2005) to equalize the expression of sex-linked genes between males (XY) and females (XX) (Lyon 1961). DNA methylation plays an important role in these processes maintaining one of the X active (Xa) and the other inactive (Xi). Some studies have shown that CpG islands have a tendency to be methylated on the Xi and unmethylated on the Xa (Tribioli et al. 1992; Hellman & Chess 2007; Ibragimova et al. 2014; Sharp et al. 2011). In the figures, it is also observed that *β*-values follow a beta distribution in the autosomal chromosomes while the *M*-values follow a bimodal distribution that accomplishes the homoscedastic characteristic. In both cases the majority of the values are in the low category as shown in the Tables 2.2.1 and 2.2.2. Low values of methylation were observed in 73% of the *β*-values and the 68% of the *M*-values. In the case of the X-chromosomes the distribution is approximated to a normal distribution having medium values of methylation. For this thesis, the *M*-values are used as the measured of DNA methylation.

Table 2.2.1. Distribution of *β*-values by autosomal chromosomes both sexes and X-chromosome in females classified by low, medium or high methylation.

	Low	Medium	High
	($\beta < 0.3$)	($\beta \in 0.3$-0.7)	($\beta > 0.7$)
Autosomal Chromosomes			
N	969,888	180,240	181,632
Freq. (%)	73%	13%	14%
X – Chromosome in females			
N	616	1,479	419
Freq. (%)	24%	59%	17%

Table 2.2.2. Distribution of *M*-values by autosomal chromosomes both sexes and X-chromosome in females classified by low, medium or high methylation.

	Low	Medium	High
	($M < -2$)	($M \in -2,2$)	($M > 2$)
Autosomal Chromosomes			
N	907,730	289,295	134,735
Freq. (%)	68%	22%	10%
X – Chromosome in females			
N	393	1,847	274
Freq. (%)	16%	73%	11%

Chapter 3: Transcriptomics from tumor tissue

This chapter contains the preprocessing, QC and exploratory single analysis of the transcriptomics data from the EPICURO project.

Gene expression data was obtained from 43 tumor samples with the Affymetrix DNA microarray Human Gene 1.0 ST platform with 32,321 probes. The DNA microarray is a collection of microscopic DNA spots attached to a solid surface. Each spot contains a specific DNA sequence known as probes. These are a short section of a gene or other DNA elements that are used to hybridize cDNA sample. Then, probe-target hybridization is usually detected and quantified by fluorescence that determines relative abundance of nucleic acid sequences in the target. Once the raw intensity levels are generated and stored in .CEL files, they are preprocessed using bioconductor affy package in R (Gautier et al. 2004) using the Robust Multi-Array Average (RMA) algorithm (Irizarry et al. 2003). This algorithm consist in three steps: (1) Background correction to remove local artifacts and background noise, (2) log2 transformation to make variation similar across orders of magnitude and (3) quantile normalization to adjust data for technical variation. Finally, a linear model fits to the normalized data to obtain expression measure for each probe set.

After preprocess the data, the QC was performed using Bioconductor arrayQualityMetrics package in R (Kauffmann et al. 2009). This package generates a report with several figures that detects if there are problems in the arrays. Figure 2.3.1 shows the distance between two arrays. This was computed as the mean absolute difference between the data of the array. Outlier detection was performed by looking for arrays for which the sum of the distance to all other arrays was exceptionally large. The array 22, 11, 19, 34, 32, 20 and 21 cluster differ from the rest of the arrays showing exceptionally large distance from the others and therefore likely outliers.

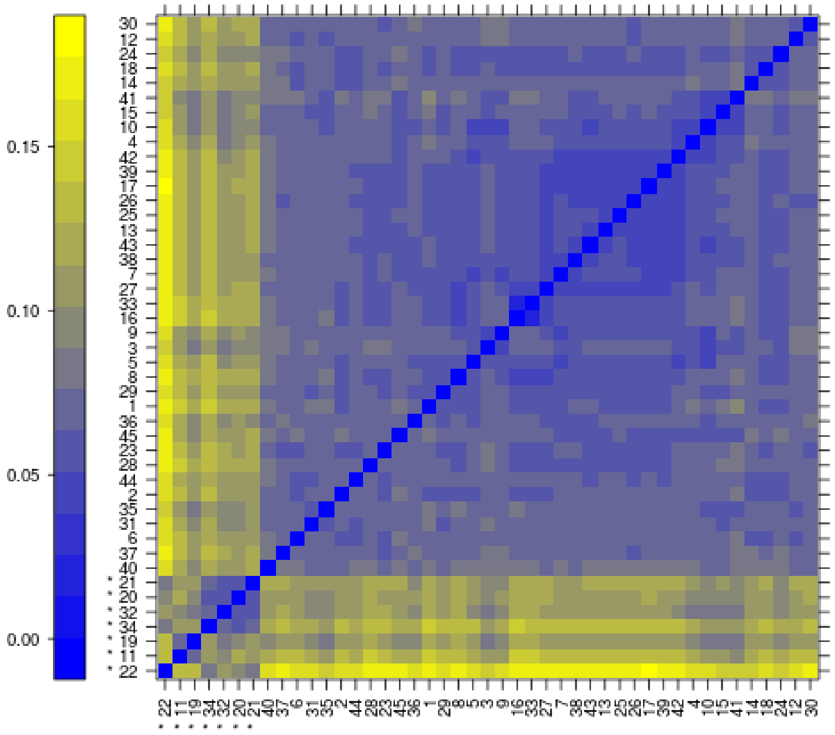


Figure 2.3.1. Distance between arrays. The color scale is chosen to cover the range of distances encountered in the dataset. Patterns in this plot can indicate clustering of the arrays either because of intended biological or unintended experimental factors (batch effects).

Figure 2.3.2 shows the distribution per sample where one expects the boxes to have similar positions and widths. When the distribution is very different from the other, this may indicate experimental problems. The detection of outliers was performed by computing the Kolmogorov-Smirnov statistic between each array's distribution and the pooled distribution. This test is applied to compare distributions and inspect whether they come from the same distribution or another. In this case, 20, 21, 22, 32 and 34 arrays were considered as outliers.

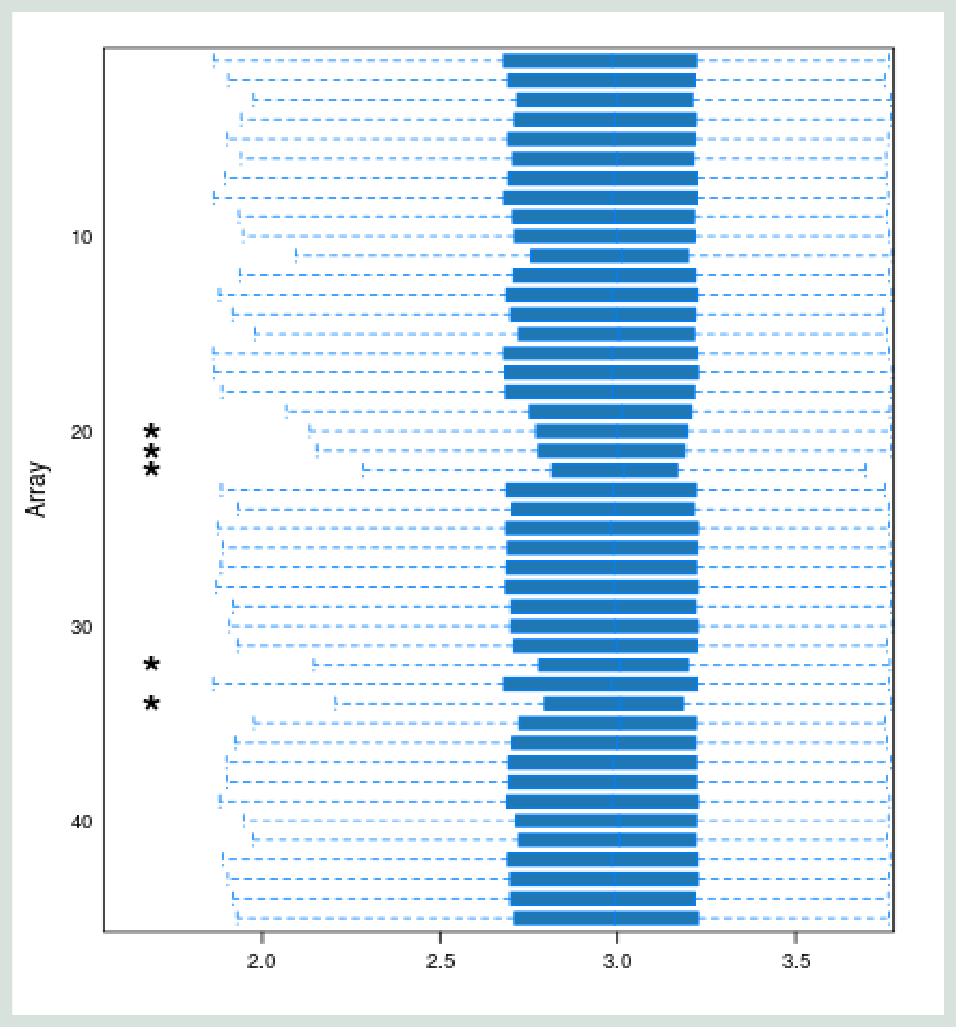


Figure 2.3.2. Boxplot representing the distribution corresponding to each array.

Finally, Figure 2.3.3 shows the mass of the distribution of M and A defined as: M = $\log_2 (I_1)$ - $\log_2 (I_2)$ and A = 1/2 ($\log_2 (I_1)$ + $\log_2 (I_2)$), where I_1 is the intensity of the array studied, and I_2 is the intensity of a "pseudo" array that consists of the median across arrays. The detection of outliers was performed by computing Hoeffding's statistic D_a on the joint distribution of A and M for each array. The figure below shows the 4 arrays with the highest value of D_a (top row) and 4 with the lowest (bottom row). This test defined an outlier when the statistic D_a > 0.15 and no outliers were marked in our data.

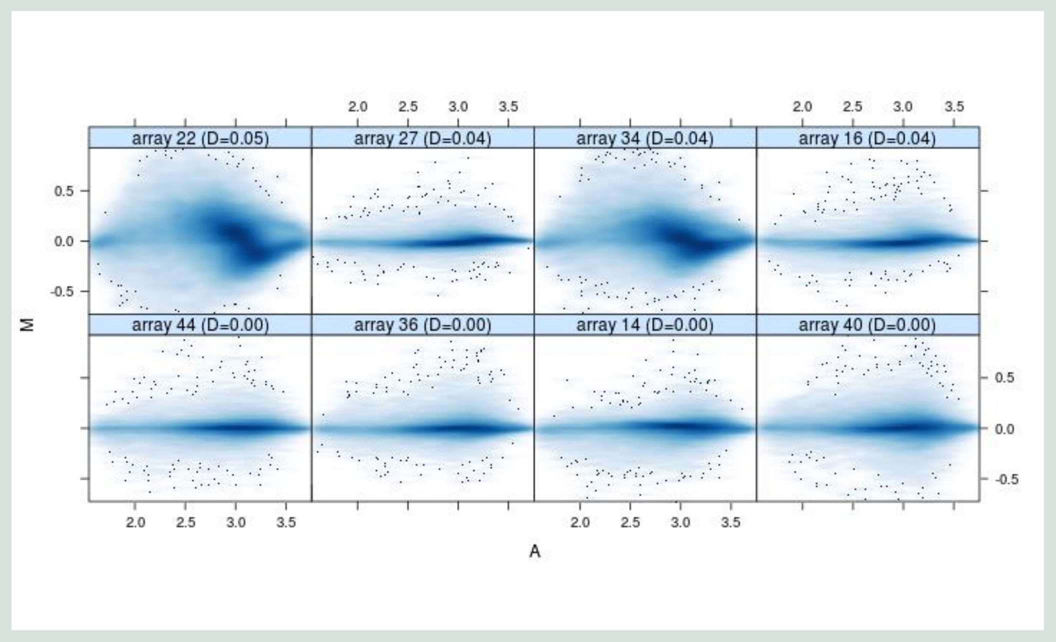


Figure 2.3.3 MA plot representing the mass of the distribution of M and A. Typically, we expect the mass of the distribution in an MA plot to be concentrated along the M = 0 axis, and there should be no trend in M as a function of A.

After the inspection of these graphs, we considered to delete the arrays marked as outlier in the Figure 2.3.1 and Figure 2.3.2 to avoid any problem in further analysis. The RMA algorithm was re-run again with the final sample set of 37 and the QC was also re-run to ensure that no outlier was detected.

After annotating the probes, we deleted the ones that were not assigned to any gene using the affymterix hugene 10 annotation data from Bioconductor in R (MacDonald JW). 20,899 probes were annotated to genes for 37 individuals and these were used in chapter 1 and 2 in PART 3 of this thesis. They were annotated using the UCSC hg19, NCBI build 37 to make them comparable and homogenize their position in the genome.

Figure 2.3.4 represents graphically the distribution of the final number of probes and samples after applying the RMA algorithm. It follows a normal distribution and therefore parametric statistics were applied to analyze these data.

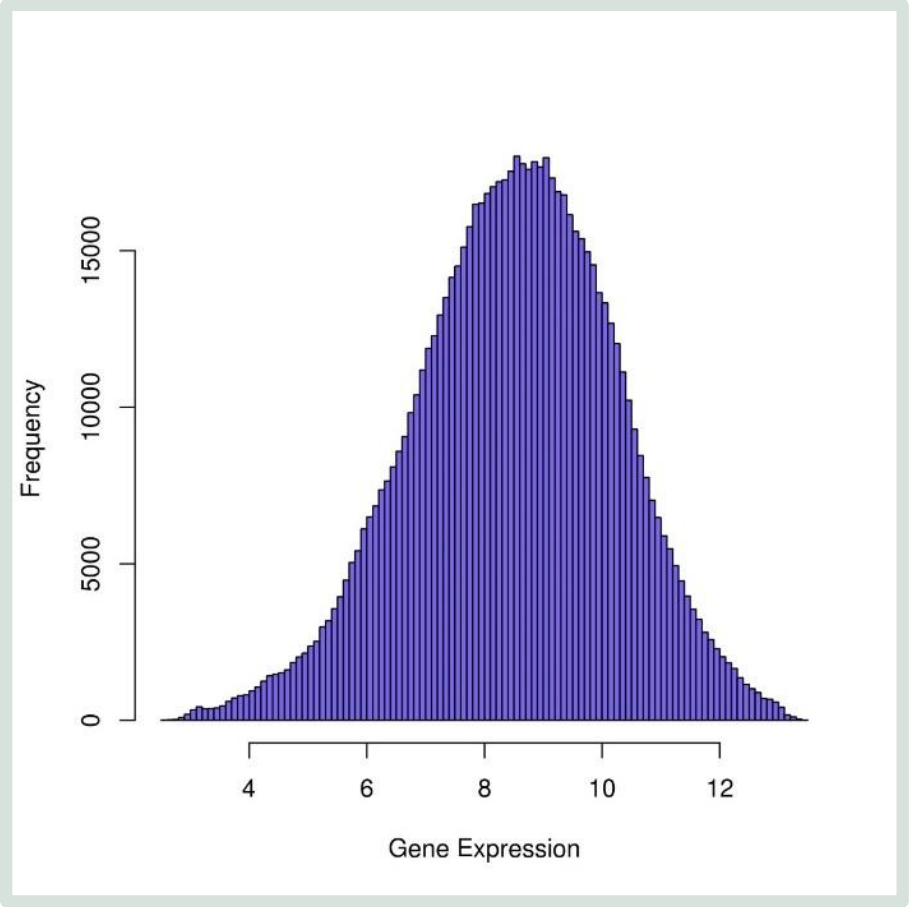


Figure 2.3.4. Distribution of gene expression data after preprocessing and QC.

PART 3.

NOVEL STATISTICAL APPROACHES FOR INTEGRATIVE –OMICS ANALYSIS

The general objective of this thesis was to dissect and fix the methodological challenges of –omics data integration where data from tumoral tissue (genomics, epigenomcis and transcriptomics) and data from blood samples (genomics) are combined. To this end, we planned three specific objectives: (1) to perform the integration in a multi-step process considering all possible pairwise combinations from tumoral samples, (2) to perform the integration in a multi-dimensional approach where all the –omics are combined together from tumoral samples and (3) to perform the integration at multi-material level using data from the different source material (tumor and blood).

In this part, we address these three specific objectives structured in three scientific papers where first, a framework to integrate the three –omics data from tumoral tissue based on pairwise combinations is proposed (published: Pineda et al. 2015 Human Heredity). Second, penalized regression methods with a permutation-based MaxT method are performed to integrate the three –omics data from tumoral tissue in the same model at the same time (accepted with minor revision: Pineda et al. 2015 PlosGenetics) and, third, an integrative eQTL –omics multi-material level is proposed using the previous approach developed (submitted: Pineda et al. 2015 AJHG).

Chapter 1. Framework for the integration of genomics, epigenomics and transcriptomics in complex diseases

Silvia Pineda(1,2), Paulina Gomez-Rubio(1), Antoni Picornell(1), Kirylo Bessonov(2), Mirai Márquez(1), Manolis Kogevinas(3), Francisco X Real (4,5), Kristel Van Steen (2,6), Núria Malats(1)

(1) Genetic and Molecular Epidemiology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.
(2) Systems and Modeling Unit, Montefiore Institute, University of Liége, Liége, Belgium
(3) Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain; Institut Municipal d'Investigació Mèdica - Hospital del Mar, Barcelona, Spain.
(4) Epithelial Carcinogenesis group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.
(5) Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain.
(6) Bioinformatics and Modeling, GIGA-R, University of Liege, Avenue de l’Hôpital 1, Liége, Belgium

Hum Hered 2015;79:124-136 (DOI:10.1159/000381184)

Abstract

Objectives: Different types of 'omics' data are becoming available in the post genome era; still a single 'omics' assessment provides limited insights to understand the biological mechanism of complex diseases. Genomics, epigenomics and transcriptomics data provide insight into the molecular dysregulation of neoplastic diseases, among them urothelial bladder cancer (UBC). Here we propose a detailed analytical framework necessary to achieve an adequate integration of the three sets of 'omics' data to ultimate identify previously hidden genetic mechanisms in UBC. *Methods*: We build a multi-staged framework to study possible pairwise combinations and integrate data in three-way relationships. SNP genotypes, CpG methylation levels, and gene expression levels were determined for a total of 70 individuals with UBC and with available fresh tumor tissue. *Results*: We suggest two main hypothesis-based scenarios for gene regulation based on the "omics" integration analysis where DNA methylation affects gene expression and genetic variants co-regulate gene expression and DNA methylation. We identified several three-way trans-association "hotspots" that are found at the molecular level and that deserve further studies. *Conclusions*: The proposed integrative framework allowed us to identify relationships at the whole genome level providing some new biological insights and highlighting the importance of integrating 'omics' data.

Introduction

Big data at the molecular field ('omics' data) is being generated at an unprecedented pace, this including genome, methylome, transcriptome, and microbiome, among others. There is a growing interest in combining the different types of 'omics' datasets that are becoming available since a single 'omics' assessment provides limited insights into the understanding of the underlying biological mechanisms of a physiological/pathological condition. For example, even when many genome-wide association studies (GWAS) have identified several Single Nucleotide Polymorphisms (SNP) involved in complex diseases, the functional implications of the susceptibility loci are still poorly understood and they only partially account for the phenotype variability. Combining different 'omics' data types seems to be a more suitable approach, as it will likely reveal previously hidden information.

The simplest form of data integration involves the combination of two different data types, common examples being genetic variants and gene expression or, more recently, genetic variants and DNA methylation. DNA methylation involves the addition of a methyl group to the 5' position of the cytosine at a Cytosine-phosphate-Guanine (CpG) site. Genomic regions with high density of CpG dinucleotides are denominated CpG islands; they are often located in gene promoters and have important roles in gene regulation. CpG sites located up to 2kb from the island's boundaries are called CpG shores and it has been demonstrated that they are also very important for gene regulation and that they are implicated in cancer (Irizarry et al. 2009). Both CpG islands and shores, when hypermethylated and located in the promoter region of a gene, negatively regulate gene repression (Jones 2012). Therefore, it is important to take into account the relationship between DNA methylation and gene regulation in order to better understand complex diseases (Portela & Esteller 2010). For example, it has been shown that hypermethylation of CpGs located in the promoter region of some tumor suppressor genes (*INK4A*, *Rb, VHL, hMLH1, BRCA1*, etc) contribute to cancer development (Esteller 2008). Therefore, analyzing gene expression data without considering epigenetics provides an incomplete genomic explanation of the transcriptome. Moreover, as DNA methylation regulates gene expression, genetic variants affecting CpG sites might, in turn, affect gene expression too. It is well known that genetic variants can alter gene expression levels and hence the importance of connecting the DNA sequence to the RNA level. The identification of these expression quantitative trait loci (eQTL) relationships may help to identify regulators of gene expression (Cheung & Spielman 2009). These eQTLs have been extensively studied to find associations between common genetic variants and gene expression levels (Nica et al. 2010; Nicolae et al. 2010; Pickrell et al. 2010; Westra et al. 2013;

Zhernakova et al. 2013). By contrast, the study of potential associations between common variants, DNA methylation levels (methylation QTLs, methQTLs), and gene expression has generated less interest, so far (Heyn et al. 2014; Gibbs et al. 2010; Zhang et al. 2010; Bell et al. 2011; Drong et al. 2013).

Genome, transcriptome, and methylome data offer unique opportunities when combined in the same analyses. This strategy has been applied to HapMap cell lines (Bell et al. 2011), whole blood from healthy human subjects (Van Eijk et al. 2012), and human monocytes (Liu et al. 2013). Furthermore, some studies have combined these types of data to better understand complex diseases, such as breast cancer (Li et al. 2013) or type 2 diabetes (Greenawalt et al. 2012). As DNA methylation is tissue-specific, these analyses have also been applied to different types of tissues, such as human brain (Gibbs et al. 2010) or adipose tissue (Drong et al. 2013)(Drong et al. 2013). It is worth noting that the majority of these studies have only assessed *cis*- relationships, but *trans*- effects deserve further study within the 'omics' context, especially as the complex organization of chromatin in the nucleus is better understood.

In the present study we built and propose a multi-staged analytical framework to integrate 'omics' data. We tested it in an urothelial bladder cancer (UBC) model using common genetic variants, DNA methylation, and gene expression transcripts data from 70 cancer patients. We proved the ability of the framework to identify some "multi-omics" relationships that provided further knowledge to better understand the biological mechanisms underlying the disease.

Material and Methods

Study Subjects: SNP genotypes, CpG methylation levels, and gene expression levels were measured for a total of 70 individuals with available fresh tumor tissue that were recruited as part of the pilot phase of the EPICURO study. All of them were histologically confirmed UBC cases recruited in 2 hospitals in Spain during 1997-1998. Tumor DNA and RNA were extracted and used for 'omics' assessment. SNP data was available for 46 patients, CpG methylation for 46 patients and gene expression for 43. The overlapping of patients between the three 'omics' was 31 for the expression-methylation relationship, 27 for the eQTL, and 46 for the methQTL studies.

SNP genotype data: Genotyping was performed using Illumina HumanHap 1M array in tumor samples. A total of 1,047,101 SNPs were genotyped in 46 individuals. For genotype calling, we used the cluster file obtained when the same array was applied to germline DNA from 2,424 subjects included in the main EPICURO study. We considered SNPs with <5% of missing values and with a minor allele frequency (MAF) ≥ 0.01. Standard Quality Control (QC) was performed using BeadStudio and R. From BeadStudio, the genotypes (AA, Aa, aa) were obtained in forward strand for those samples having a call rate higher than 90%.

DNA methylation data: After bisulphite modification of 46 tumor DNA samples using EZ-96 DNA METHYLATIONGOLD KIT (Zymo Research, Irvin, CA, USA), CpG methylation data was generated using the Infinum Human Methylation 27 BeadChip Kit that detected the CpG sites with two probes, one designed against the unmethylated site (signal U) and the other against the methylated site (signal M). The level of methylation was determined at each locus by the intensity of the two possible fluorescent signals (Bibikova et al. 2009). At each CpG site, the methylation levels were measured with the β-value, defined as:

$$\beta = \frac{\max(M,0)}{\max(U,0) + \max(M,0) + 100}$$

The maximum between signal intensity and 0 is used for β calculation to avoid the negative numbers caused by background subtractions, consequently, β-values rank between 0 (unmethylated) and 1 (methylated). The constant 100 was used to regularize the β-values when they were very small. Although β-values are useful under some circumstances, it has been demonstrated that M-values are more statistically valid than β-values due to a better approximation of the homocedasticity (Du et al. 2010). This property is important when applying regression models that require this assumption. The M-value is calculated as follows:

$$M = log_2\left(\frac{\max(M,0)+1}{\max(U,0)+1}\right)$$

It ranges between -∞ (unmethylated) and +∞ (methylated). In our study, M-values were used when applying linear regression models, while β-values were used in the rest of the analyses.

The initial number of CpGs in the studied array was 27,578. We then applied BeadStudio software and R to preprocess the data. Background normalization was performed minimizing the amount of variation in background signals between arrays and, as recommended by Illumina, CpGs were rejected when detection p-value was > 0.05. The β-values < 0 or > 1 were also excluded. CpGs with SNPs (N=908) or cross reactive probes (N=2,985) were deleted based on earlier reports for the 27K array (Chen et al. 2011). After QC, a total number of 23,034 CpGs were kept for analysis. These were classified in 3 categories for subsequent analyses: CpG islands (located in the promoter region of a gene), CpG island shores (in a sequence up to 2Kb from an island) and CpGs outside of an island or a shore.

Gene expression data: Gene expression data were obtained from 43 tumor samples using the Affymetrix DNA Microarray Human Gene 1.0 ST Array with 32,321 probes. This array was based on 2006 (UCSC hg19, NCBI build 37) human genome sequence with coverage of RefSeq, Ensembl and putative complete CDS GenBank transcripts (www.affymetrix.com). QC was performed using Bioconductor libraries in R (www.bioconductor.org/). The arrayQualityMetrics package (Kauffmann et al. 2009) was used to implement a background correction and to carry out normalization of expression levels across arrays. Application of QC steps resulted in 20,899 probes and 37 individuals. The affy library in R (Gautier et al. 2004) was used to annotate the probes.

Statistical Analysis

First, tumoral DNA methylation levels in CpG sites and gene expression levels were compared using Spearman's rank correlation for non-normally distributed variables. Second, we assessed eQTLs and methQTLs, via linear regression modeling for those expression-methylation pair probes that were strongly associated in the previous step. To perform these analyses, we obtained a linear regression model for each SNP as:

$$Gene\ Expression_i = \alpha + \beta * SNP_i$$

$$Methylation\ CpG\ _i = \alpha + \gamma * SNP_{\hat{\imath}}$$

Prior to analysis, we excluded those SNPs that had less than two individuals per genotype due to the imbalance that may produce a highly differential gene expression values, i.e: an individual with rare homozygous genotype and with an extreme gene expression value that could produce an artificial high significant p-value.

Expression-methylation probe pairs and eQTLs and methQTLs were classified in three categories according to possible genomic distance effects: *cis*-acting, if probes were located within 1Mb; *trans*-acting, if probes were on the same chromosome but located more than 1Mb apart; and *trans*-acting-outside, if they were on different chromosomes. To control the analyses for multiple testing we applied the Benjamini & Yekutieli (Benjamini & Yekutieli 2001) FDR method that allows for panel dependencies between tests. We applied this correction taking into account the number of tests performed in the eQTL and the methQTL study independently. Finally, we checked the regions of the trait-associated SNPs already published for UBC.

Third, in line with the study, we integrated the results obtained from pairwise analyses on genome, epigenome and trascriptome data. We checked the SNPs that were common in the eQTL and methQTL analysis based on those probes-CpGs that were previously correlated in order to have a complete view of the genome in individuals with UBC. We obtained the distribution of the triplets (SNP-CpG-Gene expression) that were significantly associated in the same relationship.

Statistical analyses were performed with R and results were visualized with Circos software (Krzywinski et al. 2009).

Results

The majority of the individuals included in our study were male (93%) and current (50%) or former (36%) smokers. According to established criteria based on tumor stage (T) and grade (G) for UBC, individuals were classified as having low-risk non-muscle invasive tumors (45%), high-risk non-muscle invasive tumors (22%) or muscle-invasive tumors (29%) (Table 3.1.1).

Table 3.1.1. Characteristics of the studied patients

Characteristics	N (%)
Total	72
Gender	
Male	67 (93)
Female	5 (7)
Age	
Mean (SD)	65.6 (9.5)
Min-max	41-80
Region	
Barcelona	31 (43)
Elche	41 (57)
Smoking status	
Non-smoker	8 (11)
Current	36 (50)
Former	26 (36)
Unknown	2 (3)
Tumor-stage	
Low-grade-NMIBC	32 (45%)
High-grade-NMIBC	16 (22%)
MIBC	21 (29%)
Unknown	3 (4%)

The description of the study results is organized in four sections following the framework steps proposed (Figure 3.1.1): (1) Description of the patterns of individual 'omics' data, globally and according to epidemiological data, (2) Correlation analysis between methylation and expression probes, (3) Identification of *cis*- and *trans*- eQTLs and methQTLs, and (4) Integration of results derived from the previous pairwise analysis.

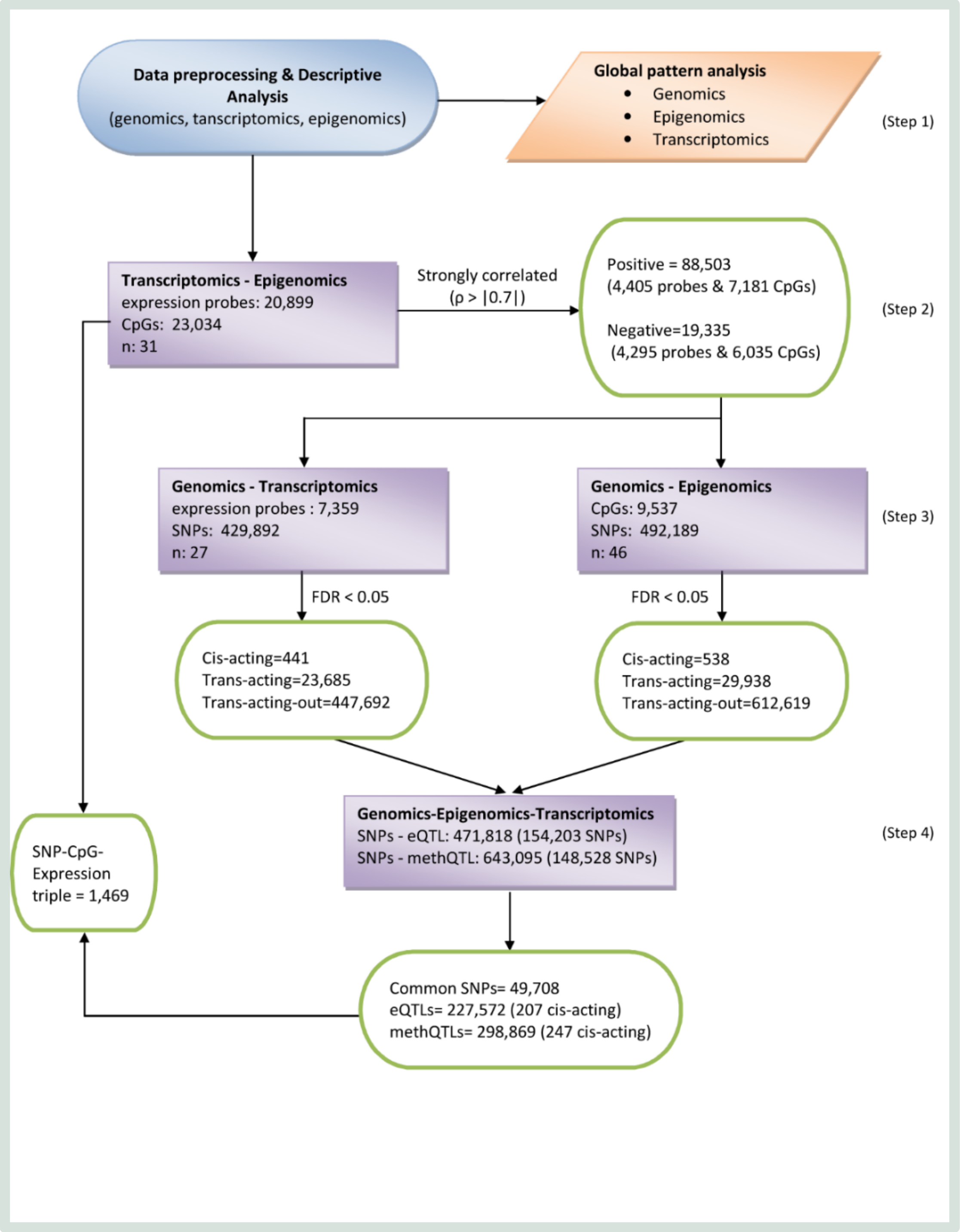


Figure 3.1.1. Framework for data integration showing the steps to integrate genetic variants, DNA methylation levels, and gene expression levels. Step 1 corresponds to the preprocessed data, quality control and global patterns individually per data set. Steps 2, 3 and 4 are represented for purple boxes corresponding to the analysis performed and the input data, and green oval boxes correspond to the results and the input of the next step.

1. Patterns of individual 'omics' data. Table 3.1.2 shows the distribution of the genotypes according to their MAF; 14% had a MAF of 0 and were excluded from the analysis, 11% ranged between (0.01-0.05], 30% between (0.05-0.2] and 31% between (0.2-0.4]. Missingness <5% was observed in 84% of the SNPs.

Table 3.1.2. Summary of SNPs genotyped

SNPs	N (%)
Total number	1,047,101
MAF	
[0.0]	150,548 (14)
(0.0 – 0.01]	0 (0)
(0.01 – 0.05]	108,496 (11)
(0.05 – 0.2]	312,220 (30)
(0.2 – 0.4]	327,762 (31)
(0.4 – 1.0]	148,075 (14)
Missingness	
No missing	488,288 (47)
5% missing	400,918 (38)
20% missing	147,732 (14)
> 20% missing	10,163 (1)

MAF = 0.0 means that all individuals are common homozygous for the measured SNP.

The patterns for DNA methylation according to the β- and M-values were different for autosomal chromosomes and X-chromosomes in females due to the X-chromosome inactivation in females. The majority (71%) of CpGs in autosomal chromosomes were unmethylated (β < 0.3) while, as expected, the majority of the CpGs (66%) in the X-chromosomes showed β-values in the range (0.3 ≤ β < 0.7). While the M-values for autosomal chromosomes displayed a bimodal distribution, X-chromosomes approximated a normal distribution (Figure 2.2.2). No significant different methylation patterns were found according to the clinical/epidemiological data considered, i.e. smoking status, tumor stage, age, and sex (Pearson's χ^2-test, data not shown).

The expression of the gene probes after background correction and normalization followed a normal distribution (Figure 2.3.4). We did not find any significant difference according to the clinical/epidemiological data by applying student's *t*-test (data not shown).

2. Correlation between gene expression and DNA methylation. While it is well established that DNA methylation may affect the expression of a gene, mainly when the relationship is in *cis*-, little is known when it is in *trans*-. We investigated a total of 481,387,566 possible correlations between gene expression and methylation both in *cis*- and in *trans*-. The number of comparisons performed was based on data derived from 31 individuals (Table 3.1.3). We obtained 19,335 strong-negative (ρ < -0.7) and 88,503 strong-positive (ρ > 0.7) associations between gene expression and methylation corresponding to 7,359 expression traits and 9,537 CpG sites. The distribution of the stronger relationships according to the CpG location and direction is shown in Table 3.1.4: 5,414 (28%) were located in CpG islands, 1,690 (59%) in CpG shores and 2,433 (57%) outside of CpG islands/shores. There were 263 (0.03%) *cis*-acting correlations, 6,177 (0.02%) *trans*-acting correlations within the same chromosome, and 101,398 (0.02%) *trans*-acting outside the chromosome (*trans-out* correlations). A whole list of CpGs with significant *cis*- association with a gene can be found in Table S3.1.1.

Table 3.1.3. Strength of correlations between gene expression and DNA methylation

Spearman's rho	Strength of correlation	Nº of combinations
(-0.9 : -1.0]	Very Strong-negative	0
(-0.7 : -0.9]	Strong-negative	19,335
(-0.4 : -0.7]	Moderate-negative	9,266,544
(-0.0 : -0.4]	Weak-negative	238,601,864
[0.0]	No correlation	380,834
(0.0 : 0.4]	Weak-positive	223,165,638
(0.4 : 0.7]	Moderate-positive	9,864,848
(0.7 : 0.9]	Strong-positive	88,503
(0.9 : 1.0]	Very Strong-positive	0

Table 3.1.4. Strong correlation for *cis*-acting and *trans*-relationships between CpG methylation and gene expression

		Negative correlation N (%)	Positive correlation N (%)
Cis-acting (same gene)	CpG island/shore	37 (80)	9 (20)
	CpG outside	3 (37)	5 (63)
Cis-acting (dif. gene)	CpG island/shore	41 (26)	116 (74)
	CpG outside	11 (21)	41 (79)
Trans-acting	CpG island/shoe	757 (17)	3,736 (83)
	CpG outside	412(24)	1,272 (76)
Trans-acting-outside chromosome	CpG island/shore	11,860 (16)	63,054 (84)
	CpG outside	6,214 (23)	20,270 (76)

3. Identification of *cis*- and *trans*- eQTLs and methQTLs. In order to detect genetic variants affecting gene expression or DNA methylation, we investigated a total of 7,359 expression traits and 9,537 CpG sites that were strongly correlated in the previous step. The number of SNPs considered here after QC was 429,892 for the eQTL and 492,189 for the methQTL analyses, resulting in a total of 3,163,575,228 eQTLs in 27 individuals and 4,694,006,493 methQTLs explored in 46 individuals. After correction for multiple testing (FDR<0.05), we obtained 471,818 significant eQTLs involving 154,203 SNPs, and 643,095 methQTLs involving 148,528 SNPs. These results pointed to the fact that multiple expression probes and CpGs were significantly associated with more than one SNP. We refer to this phenomenon as "hotspots" (Figure S3.1.1). We show the distribution of QTLs classified by genomic distance and MAF of the relationship for eQTLs in Table 3.1.5 and methQTLs in Table 3.1.6. When classifying the QTLs by genomic distance we observed 441 *cis*-eQTLs (0.02%), 23,685 *trans*-eQTLs (0.01%) and 447,692 *trans-out*-eQTLs (0.01%); and 538 *cis*-methQTLs (0.01%), 29,938 *trans*-methQTLs (0.01%), and 612,619 *trans-out*-methQTLs (0.01%). When classifying the QTLs in terms of MAF the majority had a MAF ≤ 0.2 (0.006%), while 0.003% and 0.002% had MAFs of (0.2-0.4] and ≥ 0.4, respectively. Detailed information regarding the *cis*- relationship is provided in Tables S3.1.2 and S3.1.3. When we checked how the significant findings are distributed in terms of the direction of the relationship, there were more QTLs positively than negatively (60% vs. 40% eQTL, 63% vs. 37% methQTLs) associated implying that having more copies of the rare allele increases the levels of the gene expression or the levels of methylation. Lastly, we investigated, for QTL associations in our study, how many of the SNPs involved have been previously reported as a trait associated SNPs for UBC. We found that the SNP rs401681-*TERT/CLPTM1L* on chromosome 5 was associated with the expression of *FRMD6* located on chromosome 14 (p-value = 3.7*10^{-5}), and with the cg18368125-*TMED6* on chromosome 16 (p-value = 4.8*10^{-5}). Also, the SNP rs1495741-*NAT2* on chromosome 8 was associated with the expression of *C19orf73* located in chromosome 19 (Figure 3.1.2).

Table 3.1.5: Significant (FDR<0.05) *cis*-eQTLs and *trans*-eQTLs by MAF and sign of the association

MAF	Sign	cis-eQTL N (%)	trans-eQTL N (%)	Trans-out-eQTL N (%)
(0.01-0.2]	Positive	106 (0.005)	7,026 (0.005)	127,177 (0.004)
	Negative	56 (0.002)	2,857 (0.002)	61,134 (0.002)
(0.2-0.4]	Positive	95 (0.003)	4,759 (0.003)	88,213 (0.003)
	Negative	66 (0.002)	3,220 (0.002)	65,457 (0.002)
> 0.4	Positive	57 (0.003)	2,930 (0.002)	54,087 (0.002)
	Negative	61 (0.003)	2,893 (0.002)	51,624 (0.002)

%: Percentage of significant eQTLs after multiple testing correction over the total number of *cis*- (2,331,808), *trans*- (151,738,928) and *trans*-out (3,009,504,492) eQTL

Table 3.1.6: Significant (FDR<0.05) *cis*-methQTLs and *trans*-methQTLs by MAF and sign

MAF	Sign	cis-methQTL N (%)	trans-methQTL N (%)	trans-methQTL-out N (%)
(0.01-0.2]	Positive	137 (0.004)	8,576 (0.004)	190,221 (0.004)
	Negative	61 (0.002)	3,554 (0.002)	72,611 (0.002)
(0.2-0.4]	Positive	118 (0.003)	6,864 (0.003)	139,830 (0.003)
	Negative	139 (0.004)	5,230 (0.002)	98,068 (0.002)
> 0.4	Positive	39 (0.001)	3,090 (0.001)	57,476 (0.001)
	Negative	44 (0.001)	2,624 (0.001)	54,413 (0.001)

%: Percentage of significant methQTLs after multiple testing correction over the total number of *cis*- (3,499,636), *trans*- (224,328,090) and *trans*-out (4,466,178,767) methQTL.

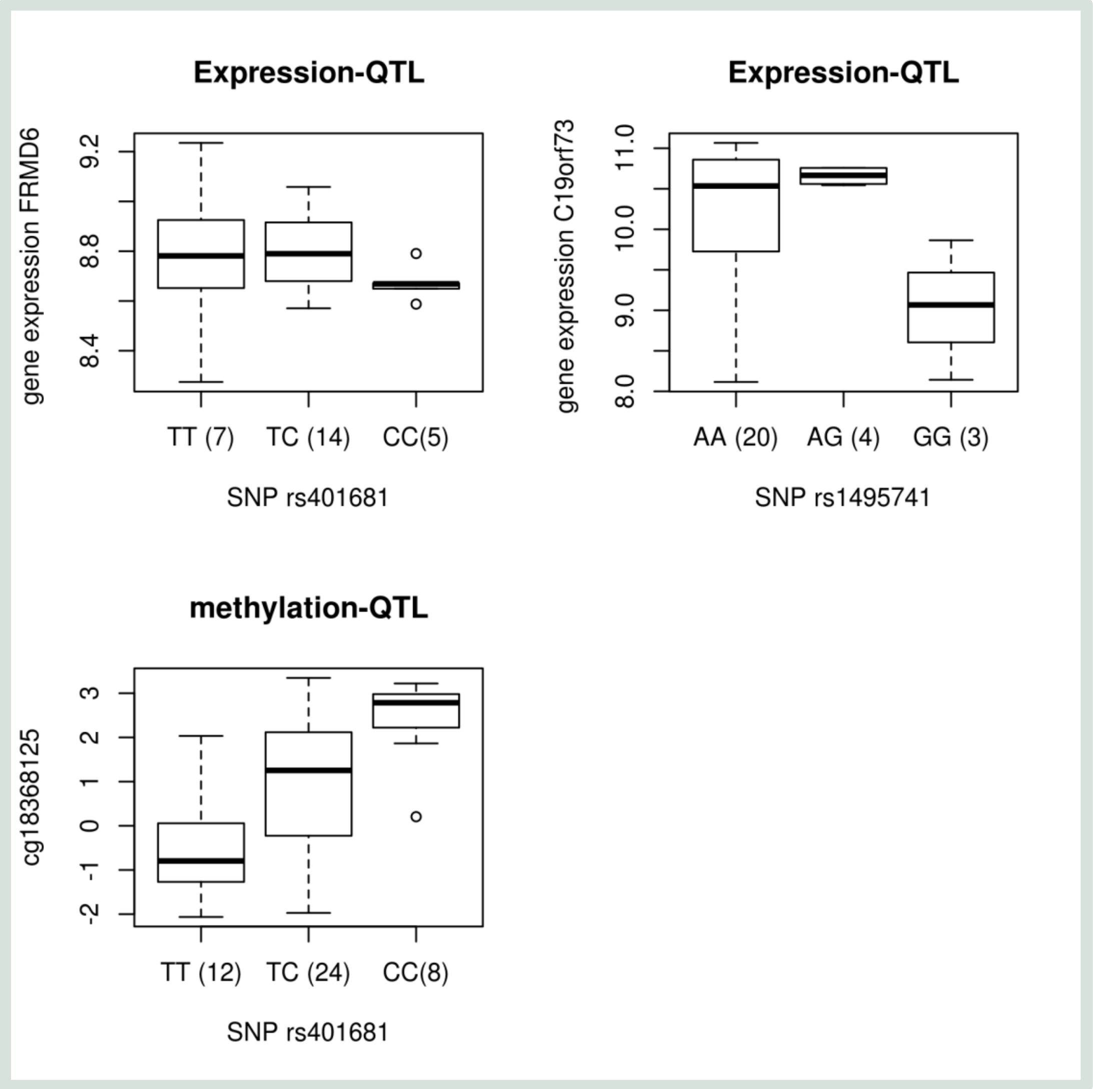


Figure 3.1.2. GWAS-reported SNPs significantly associated with gene expression levels and/or DNA methylation levels in UBC.

4. Integration of results derived from the pairwise analysis. From the final subset of eQTLs and methQTLs, we obtained 49,708 common SNPs (50% from the total SNPs for eQTLs and methQTLs), affecting a total of 227,572 eQTLs (207 *cis*-acting) and 298,869 methQTLs (247 *cis*-acting). Multiple expression probes and CpGs were significantly associated with more than one SNP and vice versa. We found that 1,469 QTLs belonged to a triple relationship (SNP-CpG-Gene expression) (Table S3.1.4). Regarding the association patterns, majority (29%) of these 1,469 triplets show a positive association pattern, that is, the higher the methylation the higher the expression, where the rare allele is classified with higher expression and methylation levels. A second pattern (19%) regarded to "the higher the methylation the lower the expression", where the rare allele is associated with high expression levels and low methylation levels. When restricted to *cis*-relationship, no triplets were found but there were 19 pairs (1 eQTL, 1 methQTL and 17 CpG-Gene expression pairs) that were in *cis*. The distribution of these triplets was completely different than that of the rest of the triplets. The most frequent pattern (32%) show a positive association between the SNP and methylation and negative for the association of both (SNPs and CpGs) with the expression. All the possible patterns with their percentages are shown in Table 3.1.7. Lastly, we checked for the “hotspots” in these triplets and we found some of them for SNPs, CpGs and Gene Expression probes (Figure 3.1.3).

Table 3.1.7: Distribution of the 1,946 triple relationships directions per pairwise analysis

eQTL	methQTL	Expr-methy	N[1] (%)	N[2] (%)
+	+	+	419 (29)	1 (5)
-	-	-	58 (4)	3 (16)
+	-	-	276 (19)	4 (21)
-	+	+	78 (5)	1 (5)
-	+	-	262 (18)	6 (32)
+	-	+	62 (4)	3 (16)
-	-	+	250 (17)	1 (5)
+	+	-	64 (4)	0 (0)

[1] The total distribution for the 1,469 triplets

[2] The distribution only for the ones that had one pair in *cis*-effect

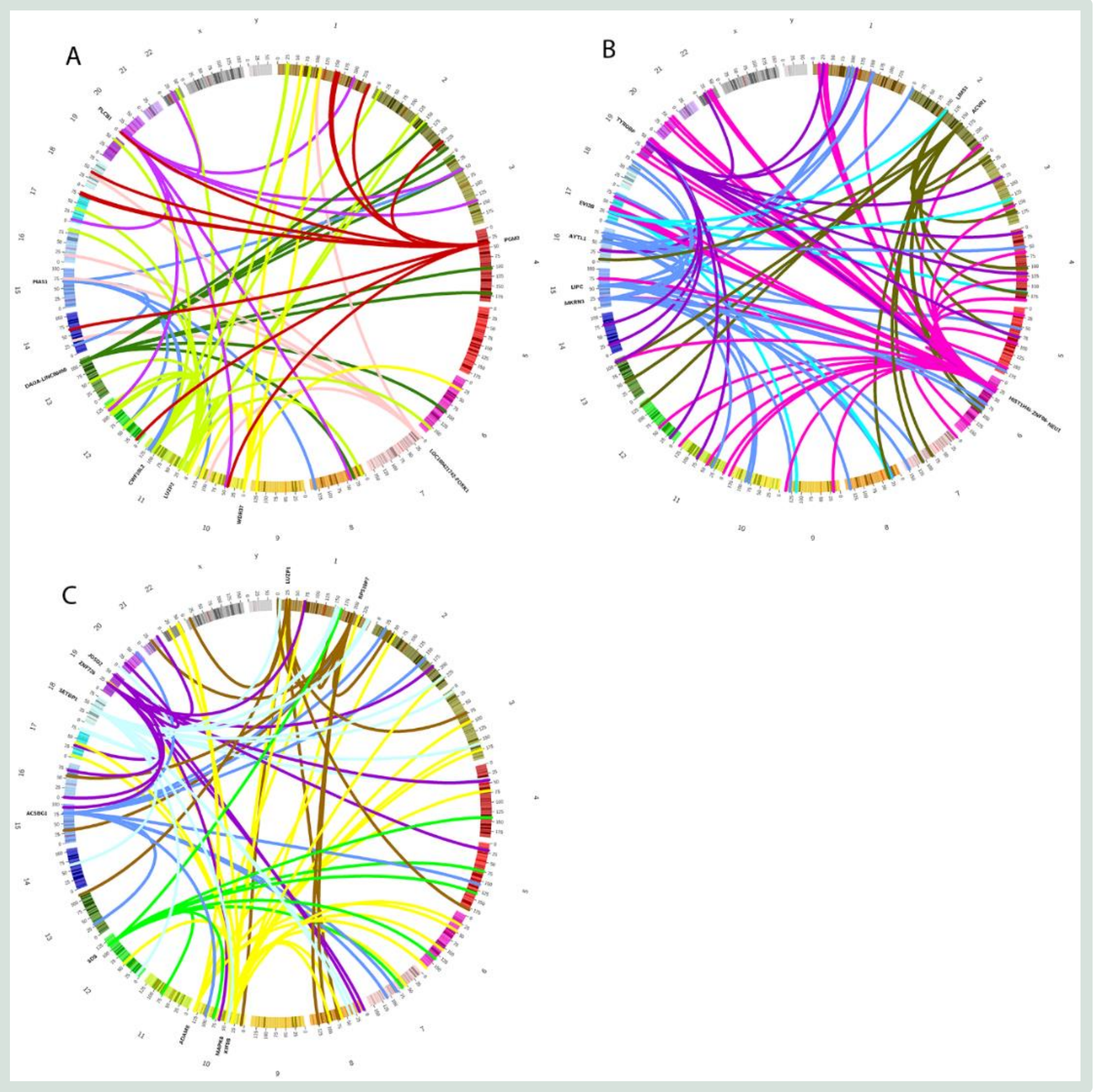


Figure 3.1.3. Circular representation of the "hotspots" found for SNPs (A), CpGs (B) and gene expression probes (C) extracted from the relationships on the triplets. Each chromosome is represented with a different color and the color of the lines corresponds to the SNPs, CpGs or gene expression probes that are located in the chromosome that share the color with. The name of the genes is located in the gene with the "hotspot".

Discussion

The post genome era delivers a wealth of ‘omics’ data allowing to explore the relationships between genetics, epigenetics and gene expression being of great importance to better understand the biological mechanism underlying a disease. In the cancer field, this integrative approach becomes particularly crucial on the basis of the knowledge indicating that SNPs, CpGs, and gene expression play an important role in the development of these complex diseases (You & Jones 2012; Kanwal & Gupta 2012).

In this work, we propose an ’omics’ integrative analytical framework based on a multi-staged strategy and we apply it to explore the relationships between three sets of data measured at a genome-wide level in UBC tumor samples. We provide further evidences on how common genetic variation and DNA methylation are statistically associated with the regulation of gene expression. Based on the knowledge that DNA is looped, allowing the interaction between two DNA regions located far away from each other, we not only studied *cis*- but also *trans*-relationships (Bickmore & van Steensel 2013). Here, we show that some SNPs are associated with DNA methylation, that the latter is associated with gene expression, and that some SNPs associate with both DNA methylation and gene expression.

Individual and pairwise analysis:

The global pattern for methylation observed in our study (Figure 2.2.2) parallels that reported previously for germline (blood) (Bell et al. 2011). Consistently with previous studies performed in blood (Bell et al. 2011; Van Eijk et al. 2012) and human brain samples (Zhang et al. 2010), we found that - when located in an island/shore - the correlations between DNA methylation and gene expression from the same gene are predominantly negative, supporting the known biological mechanisms of gene regulation (80%). DNA methylation occurs near the Transcription Start Site (TSS) of a gene, blocking the initiation of gene expression (Review in (Jones 2012)). To highlight relevant results, four different CpGs (cg01354473, cg07778029, cg25047280, cg26521404) located in a CpG island of *HOXA9* gene on chromosome 8 were negatively correlated with the expression of the gene. It was reported that *HOXA9* acts as a tumor suppressor gene in oral cancer (Uchida et al. 2014) while methylation of this gene has been associated with the regulation of its expression in UBC (Reinert et al. 2011) and with risk of different cancers such as breast (Gilbert et al. 2010), oral cavity (Guerrero-Preston et al. 2011), and ovarian (Wu et al. 2007), as well as with risk of recurrence in UBC (Reinert et al. 2012). The observed negative association between four CpGs and *HOXA9* expression in our study suggests that the inhibition of *HOXA9* expression may affect the development of UBC and supports the approach applied in this study.

On the other hand, the ENCODE Project provided some clues in the understanding of the biological behavior of *trans*- relationships and of the CpGs belonging to *cis*-relationships when located in a different gene (Encode Project Consortium 2004). In our study, we mainly observed positive correlations (79%) in all of these scenarios, meaning that increasing levels of methylation correlates with increasing levels of gene expression or the other way around, suggesting either a direct mechanism or an indirect mechanism where methylation affects expression of a gene repressor, thus leading to apparent association with increased gene levels. These results warrant further mechanistic studies explaining the complex association between DNA methylation and gene expression.

Little is known about the relationship between genetic variants and DNA methylation. Heyn et al. (2014) recently published a methQTL analysis using the cancer genome atlas data but only with SNPs detected in GWAS studies and *cis*-acting methQTLs. They detected one methQTL in UBC where the SNP rs401681 in *TERT_CLPTM1L* was associated with cg06550200 located in *CLPTM1L*; unfortunately we have not been able to replicate this association as this CpG is not present in the 27K methylation array. Nonetheless, for the first time we have performed *cis*- and *trans*- acting methQTL analysis in UBC tumor tissue samples using CpGs that were previously correlated with gene expression. From this assessment, we found 538 *cis*-relationships listed in the Table S3.1.3 with all necessary information for further studies and validation. More frequently, *cis*- relationships between genetic variants and gene expression levels have been assessed. We also performed eQTL association studies in *cis*- and *trans*- in the same conditions that for methQTLs and found 441 *cis*-eQTLs (Table S3.1.2). We performed these analyses on significant expression-methylation correlated probes identified in the first step upon the assumption that epigenetics interferes with the gene expression levels.

The proportion of eQTLs (0.01%, 471,818) and methQTLs (0.01%, 643,477) was similar, although more SNPs were involved in eQTLs (32.6%, 154,203) than in methQTLs (22.7%, 148,528), possibly because of the smaller sample size of the former. Similarly, we found no major differences in the percentages of QTL associations classified as *cis-, trans-* and *trans-out* according to the genomic distance defined before. Nevertheless, when considering the MAF distribution, a higher number of QTLs were observed for SNPs with MAF ≤ 0.2. While these results should be interpreted cautiously, due to the possibility of false positives, it is worth highlighting that we found a greater number of positive than negative QTLs relationships, meaning that having the rare allele is associated with increased gene expression or methylation levels.

Some studies have related SNPs associated with complex diseases at genome-wide significance level to gene expression or methylation levels (Heyn et al. 2014; Westra et al. 2013; Fu et al. 2012). Out of the 14 GWAS UBC SNPs (N Rothman et al. 2010), two showed to be associated with gene expression and methylation in *trans*-relationships (Figure 3.1.2). Interestingly, rs401681-*TERT/CPTL1M,* a variant strongly associated with low grade and low risk UBC (N Rothman et al. 2010), was found associated with a lower expression of *FRMD6* in our study, a gene that was reported to be involved in the inhibition of proliferation in human cells (Visser-Grieve et al. 2012).

Integrative analysis:

We observed an enrichment of significant associations of genetic variants with methylation and gene expression with 49,708 SNP related to 227,572 eQTLs and 298,869 methQTLs (207 *eQTLs* and 247 methQTL in *cis*-) suggesting a co-regulated expression and methylation. The percentage of enrichment associated with eQTLs (11.5%) and methQTLs (10.0%) was similar to that found by Wagner et al. (2014) who detected an enrichment of 9.5% in fibroblasts. Bell et al. (2011) also found an enrichment in lymphoblastoid cell lines. By contrast, Gibbs et al. (2010) found only a modest overlap between both data in brain tissues, while Drong et al. (2013) found no enrichment in adipose tissue. This highlights the fact that a specific genetic variants may show tissue-specific effects and that little is known about them at a genome wide level. We also found a total of 1,469 QTLs where the same SNP was significantly associated with both eQTL and methQTL in previously identified gene expression-CpG significant pairs. This three-way type relationship between SNP-CpG-Gene expression supports the notion that the three data sets implemented in this study are closely related in regulating part of the genome, an observation that may provide new insight into the genetics of this complex disease. Furthermore, we observed that the most frequent pattern (29%) in these three way relationships is a positive association pattern, suggesting that hypermethylation may act through a direct mechanisms or affect a repressor gene associated with an over-expression of gene levels. In addition, having the rare allele is associated with hypermethylation and over-expression pattern. This finding together with the fact that, in our study, we have demonstrated that 82% of the CpGs that are related with gene expression in *trans*-effect are positively correlated suggest that if one SNP is co-regulating both, this relation should be positive. Thus, we could hypothesize that the rare allele of the SNP associates with hypermethylation that, at the same time, associates with over-expression, as a possible regulation scenario in *trans*-effect. When inspecting the *cis*-relationships, no triplets were found, but there were 19 pairs (1 eQT, 1 methQTL and 17 CpG-gene expression pairs) that were in *cis*. In this scenario, the most frequent pattern (32%) suggests that having the rare allele is associated with hypermethylation and under-expression where the expression and methylation are associated inversely. This fact suggests another possible regulation scenario based on previous findings. We demonstrated that the 79% of the CpGs located in the promoter region of the gene are negatively correlated in *cis* with the gene expression levels; meaning that higher methylation levels may affect to a decrease in the gene expression levels. An example of this scenario is shown in Figure 3.1.4 where the SNP rs289516 located in gene *DLC1* is negatively associated in *trans* with the expression of *HOXA9* (β = -1.1; p-value = 3.7*10^{-5}) and positively with the cg01354473 located in the island of the *HOXA9* gene (β = 1.8; p-value = 9.9*10^{-5}). The relationship between the expression and the methylation levels in *HOXA9* gene was already reported as negatively correlated (r^{2} = -0.7; p-value = 1.4*10^{-5}). It has been already published that the methylation of *HOXA9* is negatively correlated with the gene expression in UBC (Reinert et al. 2011) as we observed in our study. We added a new step on this complex scenario, since the SNP rs289516 is also involved in this triple relationship. This SNP belongs to the *DLC1* gene considered as a tumor suppressor gene and the particular SNP has been picked up in two GWAS, one for asthma (Moffatt et al. 2010) and one for breast cancer (Hunter et al. 2007), but any of them passed the GWAS significant threshold. Other examples with biological support are the triplet composed by the SNP rs29658399 located in gene *DNAH11*, the gene expression of *HSPA1A*, and the cg00929855 located in gene *HSPA1A*. It has been published that the *HSPA1A* promoter methylation underlies the defect in gene expression reduction observed in UBC cell lines (Qi et al. 2013). In addition we found some "hotspots" in these triplets regarding SNPs, CpGs and gene expressions probes. In the circos plot (Figure 3.1.3 A) we observed a predominant relation for one SNP (rs10569 located in the gene *PGM2*) in chromosome 4. *PGM2* is a protein-coding gene and is associated with diseases such as pneumonia and hypoxia. While alterations in this gene have not yet been directly associated with cancer, hypoxia is a known relevant process for tumor survival. This SNP was positively associated with the expression of *SETBP1,* coding for an important cancer gene located in chromosome 18 that is observed also as a predominant "hotspot" in Figure 3.1.3 C. Somatic mutations in *SETBP1* (Piazza et al. 2013), as well as its expression patterns (Makishima et al. 2013), are related with myeloid leukemia disease. Moreover in Figure 3.1.3 B we observed a very predominant "hotspot" regarding three CpGs belonging to three different genes but close located in chromosome 6; Two of them (cg02622316 located in the gene

ZNF96 and cg02599464 located in the gene *HIST1H41)* were already published as hypermethylated in individuals with muscle invasive bladder cancer (Ibragimova et al. 2014). The first one is associated positively with many SNPs and gene expression probes and the second is associated positive and negative with some SNPs and positively with some gene expression probes. A more detailed discussion of the potential biological findings than involved the triple relationships is beyond this particularly study and detailed results about all the combinations are provided in Table S3.1.4.

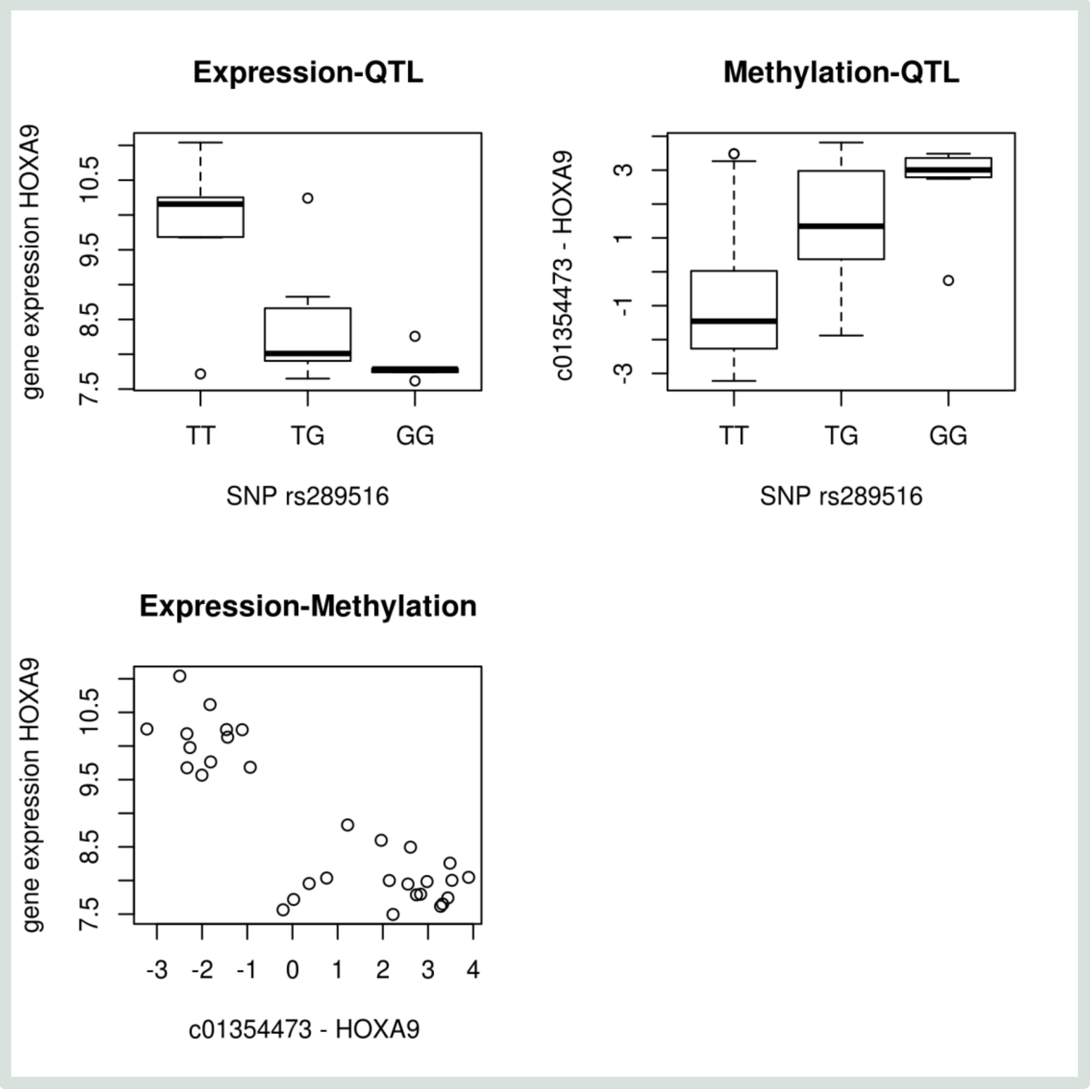


Figure 3.1.4. Example of one triple relationship where integrated common genetic variants with DNA methylation and gene expression in one of the main possible scenarios for regulation.

The integrative framework:

We built and propose a multi-staged 'omics' integration framework that its application does not require a strong methodological knowledge, being easy and effective to use. The multi-staged framework we applied has the advantage of analyzing data of all subjects that overlap among pairs of data and has not to restrict only to the few individuals with a complete overlap among all the data types. Thus, we take advantage of more samples using this framework than integrating the data in a multi-dimensional model. Therefore, we show here the application for the first time of multi-staged framework that allowed us to (1) integrate more than two 'omics' data for the same set of individuals, (2) dissect the biological relationships that may point to new mechanisms involved in the development/progression of UBC through a hypothesis-based models built step by step, and (3) to envision the complexities of the general scenario of genomic regulation.

Conclusions:

While these results are exciting, we acknowledge the following limitations. First, in this study we use the 27K methylation array that only covers a selection of CpG sites making infeasible to replicate previous reported findings using the 450k array. Second, statistical power is a commonplace in any QTL analysis given the extensive amount of data analyzed and the small sample size. While this limitation needs to be considered in the interpretation of the results, it is worth mentioning that a large enough size will unlikely be available to meet the standard criteria of statistical power; therefore, our study represents a proof of concept in the integrative 'omics' field. In addition, while we might not be able to address for unmeasured confounding factors, no differences were found between demographic factors and methylation and gene expression in our series. Validation of these results to discard false positive findings is not trivial due to the multiple genomic factors, the models considered, and the characteristics of the series. Despite these limitations, this study has several strengths. We have performed the study in tumor samples what gave us the opportunity to study in detail the regulation of three types of 'omics' data in UBC providing some evidences on the genomics regulation of the tumor. We have applied an easy, reproducible, and detailed framework to perform an integrative study of the relationships between genetic variations, DNA methylation and gene expression, showing a whole spectrum of the associations between them. We have shown that 'omics' data integration helps unraveling biological mechanisms involved in UBC. All these relations may help in the identification of new molecular targets to be further explored in detail, mainly regarding *trans*- relationships.

In conclusion, this study provides the scientific community with a pipeline to integrate more than two sets of ‘omics’ data that can be applied in future analyses seeking to better understand the biology behind the complex diseases. In addition, we highlight the importance of integrating ‘omics’ data to identify new genetic mechanisms in UBC. While several pieces of evidences support these findings, they still require of experimental validation to be considered conclusive.

Chapter 2. Integration analysis of three –omics data using penalized regression methods: An application to bladder cancer

Silvia Pineda (1,2), Francisco X Real (3), Manolis Kogevinas (4), Alfredo Carrato A (5), Stephen J. Chanock (6), Núria Malats* (1), Kristel Van Steen* (2,7)

(1) Genetic and Molecular Epidemiology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.
(2) Systems and Modeling Unit – BIO3, Montefiore Institute, Liège, Belgium.
(3) Epithelial Carcinogenesis Group, Spanish National Cancer Research Centre (CNIO), Madrid, and Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain.
(4) Centre for Research in Environmental Epidemiology (CREAL) and Parc de Salut Mar, Barcelona, Spain.
(5) Servicio de Oncología, Hospital Universitario Ramon y Cajal, Madrid, and Servicio de Oncología, Hospital Universitario de Elche, Spain
(6) Division of Cancer Epidemiology and Genetics, National Cancer Institute, Department of Health and Human Services, Bethesda, Maryland, USA.
(7) Systems Biology and Chemical Biology, GIGA-R, Liège Belgium.
* Equal contributions

Plos Genetics 2015 – (minor revision)

Abstract

Omics data integration is becoming necessary to investigate the genomic mechanisms involved in complex diseases. During the integration process, many challenges arise such as data heterogeneity, the smaller number of individuals in comparison to the number of parameters, multicollinearity, and interpretation and validation of results due to their complexity and lack of knowledge about biological processes. To overcome some of these issues, innovative statistical approaches are being developed. In this work, we propose a permutation-based method to concomitantly assess significance and correct by multiple testing with the MaxT algorithm. This was applied with penalized regression methods (LASSO and ENET) when exploring relationships between common genetic variants, DNA methylation and gene expression measured in bladder tumor samples. The overall analysis flow consisted of three steps: (1) SNPs/CpGs were selected per each gene probe within 1Mb window upstream and downstream the gene; (2) LASSO and ENET were applied to assess the association between each expression probe and the selected SNPs/CpGs in three multivariable models (SNP, CPG, and Global models, the latter integrating SNPs and CPGs); and (3) the significance of each model was assessed using the permutation-based MaxT method. We identified 48 genes whose expression levels were significantly associated with both SNPs and CPGs. Importantly, 36 (75%) of them were replicated in an independent data set (TCGA) and the performance of the proposed method was checked with a simulation study. We further support our results with a biological interpretation based on an enrichment analysis. The approach we propose allows reducing computational time and is flexible and easy to implement when analyzing several types of *omics* data. Our results highlight the importance of integrating *omics* data by applying appropriate statistical strategies to discover new insights into the complex genetic mechanisms involved in disease conditions.

Author summary

At present, it is already possible to generate different type of *omics* – high throughput – data in the same individuals. However, we lack methodology to adequately combine them. Many challenges arise while the amount of data increases and we need to find the way to identify and understand the complex relationships when integrating data. In this regard, new statistical approaches are needed, such as the ones we propose and apply here to integrate three types of *omics* data (genomics, epigenomics, and transcriptomics) generated using bladder cancer tumor samples. These innovative approaches (LASSO and ENET combined with a permutation-based MaxT method) allowed us to find 48 genes whose expression levels were significantly associated with genomics and epigenomics markers. The adequacy of this approach was confirmed by the use of an independent data set from The Cancer Genome Atlas Consortium: 75% of the genes were replicated. Previous sound biological evidences further support the results obtained.

Introduction

Integrating different *omics* data types, such as genomics, epigenomics and transcriptomics, may provide a new strategy to discover unknown genomic mechanisms involved in complex diseases (Greenawalt et al. 2012; Li et al. 2013; Serizawa et al. 2011). In cancer, tumor initiation and progression are the consequence of alterations in multiple pathways and biological processes including gene mutations, epigenetic changes, modifications in gene regulation, and environmental influences. In the process to integrate all of this information many challenges arise, among them the high dimensionality of data - since >2 *omics* data sets with millions of measurements are available from the same set of individuals - and the huge heterogeneity of *omics* data due to the different measurement scales (Hamid et al. 2009). Besides that, the data might be highly correlated, i.e. Single Nucleotide Polymorphisms (SNPs) that are in high linkage disequilibrium (LD) block or DNA CpG sites that belong to the same CpG island, contributing to multicollinearity in the analysis. Another challenge in *omics* data integration regards to the very small number of individuals in comparison to the number of parameters (“n << p”). In addition, interpretation and validation of *omics* derived results require of resources that are still lacking at present. In this rapidly evolving scenario, advanced methodological techniques are continuously emerging, demanding the development of improved data analysis tools (Chadeau-Hyam et al. 2013; Kristensen et al. 2014; Ritchie et al. 2015).

Integrative *omics* analysis refers to the combination of at least two different types of *omics* data. Relationships between two sets of *omics* parameters such as the expression quantitative trait loci (eQTL) (Shpak et al. 2014; Bryois et al. 2014; Li et al. 2013) or the methylation-QTL (methQTL) (Serizawa et al. 2011; Drong et al. 2013; Heyn et al. 2014), have been recently reported. The approach most commonly used for this type of pairwise analysis has been univariate models (i.e., Spearman/Pearson correlation or linear regression models), assuming that the changes in gene expression levels are only affected by one parameter. Until present, the combination of >2 *omics* data has been less explored. Towards this end, the previously mentioned challenges are magnified and there is a lack of advanced methodologies to deal with them. Recently, we published an integrative framework as a first approach to integrate genomics, epigenomics, and transcriptomics in individuals with urothelial bladder cancer (UBC) (Pineda, Gomez-Rubio, et al. 2015). In that work, we found that some gene expressions were co-regulated by both DNA methylation and genetic variants, both acting together in *trans* relationships. Therefore, the integration of multiple types of *omics* data by applying multivariable approaches becomes essential to understand the intricacy of the genomic mechanisms behind complex diseases and to overcome the above mentioned challenges.

In this regard, previous developments are Principal Component Analysis (PCA), to reduce data dimensionality, or Canonical Correlation Analysis (CCA) to investigate the overall correlation between two sets of variables. However, these methods are descriptive or exploratory techniques rather than hypothesis-testing tools. While some statistical applications have been developed in an *omics* integrative framework (sparse canonical correlation analysis (Parkhomenko et al. 2009), multiple factor analysis (de Tayrac et al. 2009), or multivariate partial least square regression (Palermo et al. 2009)), none of them offers the possibility to combine >2 *omics* data together in the same model.

The Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Tibshirani in 1996 (Tibshirani 1996) and the Elastic Net (ENET) proposed by Hui Zou and Trevor Hastie in 2005 (Hui Zou 2005) are penalized regression methods that, after appropriate standardization, can model more than one type of *omics* data, face multicollinearity issues, and mitigate the “n << p” problem. More importantly, both methods simultaneously execute variable selection and parameter estimation, thus reducing the computation time, while the traditional methods work on the two problems separately, first selecting the relevant parameters and then computing the estimates. LASSO and ENET have already been applied to GWAS studies (Pineda et al. 2014; Cho et al. 2010; Zhou et al. 2010) as well as in the context of integrative studies (Mankoo et al. 2011). One limitation of penalized regression techniques is that the penalty produces biased estimators; consequently, standard errors are not meaningful and cannot provide p-values to assess significance. Here, we propose a permutation-based approach to assess significance and we combine it with a correction for Multiple Testing (MT) using the MaxT algorithm (Peter H. Westfall & Young 1993). We apply this permutation-based MaxT method with LASSO and ENET to identify relationships between common genetic variation, DNA methylation, and gene expression, all determined in UBC tumor samples. Specifically, we first built a two *omics* integrative model associating SNPs or CpGs with gene expression levels and, then, we integrated the three *omics* data to assess whether changes in gene expression levels could be confounded/modified by genetic variants and/or DNA methylation.

Material and Methods

Penalized Regression Methods

LASSO and ENET penalized regression methods are applied to high-dimensional problems with a large number of parameters. The penalization produces a shrinkage of the regression coefficients towards zero given a sparse model reducing the irrelevant parameters. Both methods deal with highly correlated variables though in a different way. LASSO tends to select one variable from a group of correlated features whereas ENET selects the whole group of variables, when evidence for their relevance exists. The shrunk estimators introduce a bias while reducing the variance resulting in a better precision and accuracy model and, therefore, increasing its statistical power.

Definition of the methods

Consider the standard linear regression model where $y = (y_1, \dots y_n)^t$ is the response variable and $x = (x_{1j}, \dots x_{nj})^t \; j = 1, \dots p$ are the standardized predictors, the LASSO solves the *l₁* penalized regression problem, the Ridge regression (Hoerl & Kennard 1970) solves the *l₂* penalized regression problem and the ENET is the combination between the *l₁* and *l₂* penalized regression problem.

For the LASSO and ENET estimates $\widehat{\beta_0}$, $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^t$; $(\widehat{\beta_0}, \hat{\beta})$ are defined by

$$(\widehat{\beta_0}, \hat{\beta}) = \arg min \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}$$

$with\ the\ restrictions$:

$$\sum_{j=1}^{p} |\beta_j| \leq t \; (LASSO), \qquad (1)$$

$$\sum_{j=1}^{p} |\beta_j| \leq t \;, \sum_{j=1}^{p} \beta_j^2 \leq t \quad (ENET). \qquad (2)$$

Here, $t \geq 0$ is the tuning parameter that controls the amount of shrinkage that is applied to the estimates. For $\hat{\beta}_j^0$ the un-penalized least squares estimate, $t_0 = \sum |\hat{\beta}_j^0|$. Values of $t < t_0$ will lead to shrinkage towards 0; some coefficients may be exactly equal to 0. Using the Lagrangian form, this optimization problem is equivalent to

(LASSO): (3)

$$\hat{\beta}_{lasso} = argmin\left\{\frac{1}{N}\sum_{i=1}^{N}(y_i - x_i\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

where λ is the penalty parameter related to t. To obtain the optimal penalty, k-fold cross validation (CV) was applied (Trevor Hastie; Rob Tibshirani; Jerome Friedman 2001) maximizing the penalized log-likelihood function.

(ENET):

$$\hat{\beta}_{enet} = argmin\left\{\frac{1}{N}\sum_{i=1}^{N}(y_i - x_i\beta)^2 + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2\right\}, \qquad (4)$$

where λ_1, λ_2 are the penalty parameters related to t. In this sense, ENET can be viewed as a penalized least squares method. With $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$, solving $\hat{\beta}_{enet}$ in equation (4) is equivalent to the following optimization problem:

$$\hat{\beta}_{enet} = argmin\left\{\frac{1}{N}\sum_{i=1}^{N}(y_i - x_i\beta)^2 + (1-\alpha)\sum_{j=1}^{p}|\beta_j| + \alpha\sum_{j=1}^{p}\beta_j^2\right\} \qquad (5)$$

This expression involves a convex combination of the LASSO and ridge penalty. When $\alpha = 1$ the ENET becomes ridge regression and when $\alpha = 0$ the ENET becomes LASSO. To obtain the optimal penalty (λ), k-fold CV selecting the best α was applied. This value was obtained using a vector of $\alpha\epsilon(0.01, 0.99)\ by\ 0.01$. The LASSO and ENET methods described above were applied to our data with the R package glmnet, that relies on cyclical coordinate descent, computed along a regularization path (Jerome Firedman; Trevor Hastie; Rob Tibshirani 2010). To avoid small sample size limitations in variable selection while not introducing an important bias k = 5 was used in the k-fold CV.

These methods are promising in the context of high-throughput data but one of their drawbacks is that they do not provide p-values to assess statistical significance of relationships, nor give a formal assessment of the overall goodness-of-fit. Therefore, a permutation based strategy was adopted to assess significance of discovered relationships combined with a MT correction approach (MaxT algorithm (Peter H. Westfall & Young 1993)) building upon the statistical concept of deviance. The deviance is used to compare two models and in this case we defined it as:

$$Deviance = 2[loglik(full_{model}) - loglik(null_{model})].$$

Here $loglik$ is the loglikelihood function, $full_{model}$ refers to the model with the parameters selected by LASSO or ENET, and $null_{model}$ is the model with only the intercept estimated. Thus, the interpretation would be, the higher the deviance the better the model.

Permutation-based MaxT method

MaxT algorithm of Westfall & Young (Peter H. Westfall & Young 1993) is a step-down FWER-controlling MT procedure. The method uses the raw p-values or directly the statistics as explained in (Ge et al. 2003). Using this aproach, the permutation needed to obtain the p-values was combined with the one needed to apply the MaxT algorithm saving computational time. In this work, we used the deviance obtained per each of the permuted LASSO/ENET model to compute the MaxT algorithm and individuals within gene expression measure were permuted, that is the dependent variable in the models. The algorithm is explained in Box 2.

Box 2. Permutation-based MaxT method

From the original data, order the deviance obtained per each observed statistics: $|D_{s1}| \geq |D_{s2}| \geq |D_{s3}| \geq \cdots \geq |D_{sm}|$.

For the bth permutation, b=1…B

1. Permute the n individuals of each of the vectors $Y_m = (y_1, \ldots y_n)_m$
2. Compute the statistics $D_{1b,\ldots}D_{mb}$
3. Compute the $U_{i,b} = \max_{l=i\ldots m}\left|\mathrm{D}_{sl,b}\right|$, the successive step-down procedure is: $U_{m,b} = \left|\mathrm{D}_{sm,b}\right|$

   …

   $U_{2,b} = \max\left|\mathrm{D}_{s2,b}, \mathrm{D}_{s3,b}, \ldots, \mathrm{D}_{sm,b}\right|$

   $U_{1,b} = \max\left|\mathrm{D}_{s1,b}, \mathrm{D}_{s2,b}, \mathrm{D}_{s3,b}, \ldots, \mathrm{D}_{sm,b}\right|$
4. The steps are repeated B times and the adjusted p-values are estimated by:

$$P_{adj,i} = \frac{\#\{b; U_{ib} \geq |D_{si}|\}}{B} \; for\ i = 1 \ldots m$$

Discovery phase: The Spanish Bladder Cancer/EPICURO Study

70 patients with a histologically confirmed UBC were recruited in 2 hospitals during 1997-1998 as part of the pilot phase of the Spanish Bladder Cancer/EPICURO Study. According to established criteria based on tumor stage and grade for UBC, the tumors were classified as low-grade non-muscle invasive, high-grade non-muscle invasive, and muscle invasive. Three sets of *omics* data were obtained using fresh tumor tissue, including common genetic variation (GSE51641), DNA methylation (GSE71666), and gene expression (GSE71576). The three *omics*

data overlapped in 27 individuals that are included in this study and comprise 44% low-grade non-muscle invasive tumors, 30% high-grade non-muscle invasive tumors and 26% muscle invasive tumors. S3.2.1 Table shows the IDs of the 27 samples used in the following analysis. The local ethics committee of the participating centers approved the study and written informed consent was obtained from all participants at the time of recruitment.

Genotyping of tumor samples was performed using Illumina HumanHap 1M array. A total of 1,047,101 SNPs were determined in 46 individuals and, after the standard quality control and filter the SNPs that were in perfect LD (r^2=1), they resulted in 567,513 SNPs. The application of multivariable models required no missing values, so genotypes were imputed with BEAGLE 3.0 method (Browning & Browning 2007). CpG methylation data was generated using the Infinium Human Methylation 27 BeadChip Kit. At each CpG site, the methylation levels were measured with M-values using the log2 transformation of the β-values since they are more statistically valid due to a better approximation of the homoscedasticity. The initial number of CpGs in the studied array was 27,578 and after background normalization and QC, a total number of 23,034 CpGs were left for analysis. Gene expression data were obtained from 44 tumor samples using the Affymetrix DNA Microarray Human Gene 1.0 ST Array with 32,321 probes. After the application of QC, it resulted in 20,899 probes determined in 37 individuals. Further details about the preprocessing of the data and the quality control applied can be found elsewhere(Pineda et al. 2015). The three measures were annotated using the UCSC hg19, NCBI build 37 to make them comparable and homogenize their position in the genome.

Simulation Study

To generate a simulation sample, the association between SNPs and/or CpGs with gene expression was broken and therefore no significant results should be observed. To do that, 10-gene expression probes were randomly selected from our discovery sample showing no correlation structure between the probes and following a multivariate normal distribution. Then, the mean (μ= 8.4) and variance (σ^2= 0.4) of all the probes together were obtained. Finally, a simulated set of gene expression probes was generated using the normal distribution obtained and considering the same sample size of the discovery phase (p= 20,899 probes and N= 27 individuals).

Replication phase: The Cancer Genome Atlas (TCGA)

UBC tumor data were obtained from The Cancer Genome Atlas (TCGA) consortium (https://tcga-data.nci.nih.gov/tcga/) to replicate our findings. Data was downloaded and processed with the TCGA-Assembler (Zhu et al. 2014). The study included only individuals with muscle invasive UBC and the tumors were profiled with genome wide 6.0 Affymetrix, RNASeqV2, and

HumanMethylation450K Illumina arrays yielding data for 20,502 gene expression probes, 905,422 SNPs, and 350,271 CpGs. The total number of individuals with overlapping data from the three platforms was 238 and they were used in the replication phase of this contribution. S3.2.2 Table shows the IDs corresponding to these 238 samples.

Overall analysis flow

Penalized regression methods LASSO and ENET were applied to the discovery data in combination with the proposed permutation-based MaxT method to select the SNPs and/or CpGs associated with gene expression levels in the following multivariable models:

SNP model:

$$Gene\ Expression\ levels_i = \alpha_1 SNP_1 + \alpha_2 SNP_2 + \cdots + \alpha_p SNP_p; i = 1 \dots m$$

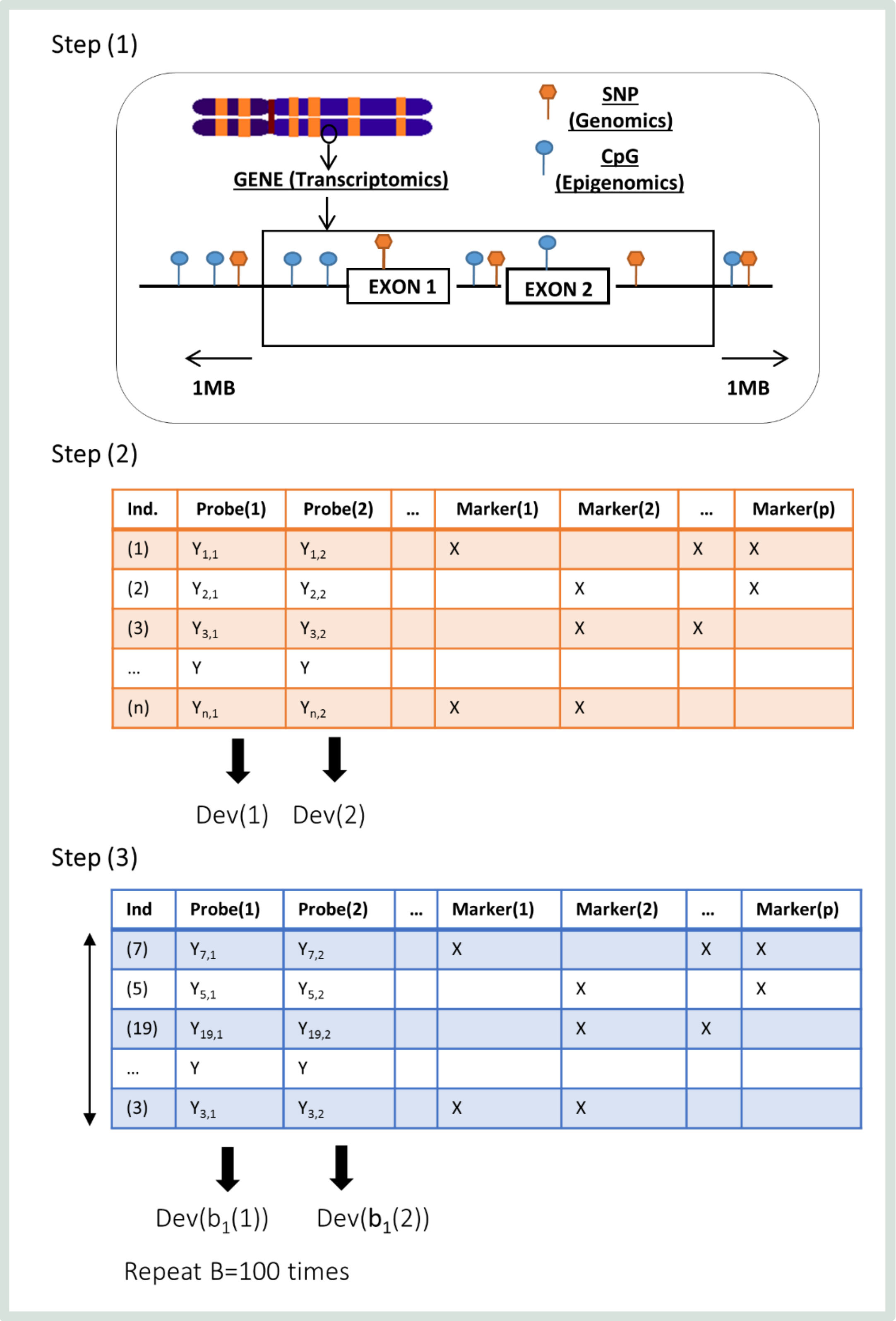
CPG model:

$$Gene\ Expression\ levels_i = \gamma_1 CPG_1 + \gamma_2 CPG_2 + \cdots + \gamma_p CPG_p; i = 1 \dots m$$

Global model = SNP + CPG model:

$$Gene\ Expression\ levels_i = \alpha_1 SNP_1 + \cdots + \alpha_p SNP_p + \gamma_1 CPG_1 + \cdots + \gamma_p CPG_p; i = 1 \dots m$$

To apply this integrative idea to our set of data the following steps were performed: (1) SNPs and CpGs that were in a 1MB window upstream and downstream were selected from each probe in the gene expression array; (2) LASSO and ENET were applied to each probe and model (SNP, CpG, and Global models) obtaining the deviance per model; and (3), the permutation-based MaxT method was applied to obtain the adjusted p-values (B= 100 permutations and significant adjusted p-value < 0.1). The scenario and workflow is represented in Figure 3.2.1.



Ind.	Probe(1)	Probe(2)	...	Marker(1)	Marker(2)	...	Marker(p)
(1)	$Y_{1,1}$	$Y_{1,2}$		X		X	X
(2)	$Y_{2,1}$	$Y_{2,2}$			X		X
(3)	$Y_{3,1}$	$Y_{3,2}$			X	X	
...	Y	Y					
(n)	$Y_{n,1}$	$Y_{n,2}$		X	X		

Ind	Probe(1)	Probe(2)	...	Marker(1)	Marker(2)	...	Marker(p)
(7)	$Y_{7,1}$	$Y_{7,2}$		X		X	X
(5)	$Y_{5,1}$	$Y_{5,2}$			X		X
(19)	$Y_{19,1}$	$Y_{19,2}$			X	X	
...	Y	Y					
(3)	$Y_{3,1}$	$Y_{3,2}$		X	X		

Figure 3.2.1. Scenario and workflow of the overall analysis implemented. The proposed integrative framework is based on three steps. Step 1 corresponds to the selection of SNPs and CpGs in 1MB window upstream and downstream from each probe in the gene expression array. Step 2 corresponds to the application of the LASSO and ENET to each probe obtaining the deviance per probe. Step 3 corresponds to the permutation-based MaxT method application where individuals are permuted B=100 times obtaining the deviance per probe.

Subsequently, this analysis flow was applied to the simulated data set using the same criteria. In the replication scenario, we aimed at determining whether the genes that were significant in the discovery phase were also significant in the replication dataset. Therefore, the analysis was restricted to the genes found to be significant in the discovery phase considering all models (SNP, CPG and/or Global) and methods (LASSO and/or ENET). Following the pipeline shown in Figure 3.2.1, we focused on the significant genes found in the discovery phase and SNPs and CpGs were selected in 1MB window from the TCGA database, even if the SNPs and CpGs were not the same as those analyzed in the discovery phase. Second, LASSO and/or ENET were conducted to SNP, CPG, and/or Global models. Finally, the permutation-based MaxT method was applied to obtain significance and correct for multiple testing. The replication analysis was performed with the same software and criteria as in the discovery analysis.

Gene enrichment analysis

To provide a biological interpretation to the results, the entire list of the significant genes identified in the discovery phase by both LASSO and ENET, and by the three models, was used to perform a gene enrichment analysis with the bioinformatics tool DAVID (Dennis et al. 2003; Huang et al. 2009). The functional annotation clustering analysis module offered by DAVID was used. The gene term annotation is based on 14 annotation categories (Gene Ontology (GO), Biological process, GO Molecular Function, GO Cellular Component, KEGG Pathways, BioCarta Pathways, Swiss-Prot Keywords, BBID Pathways, SMART Domains, NIH Genetics Association DB, UniProt Sequence Features, COG/KOG Ontology, NCBI OMIM, InterPro Domains, and PIR Super-Family Names) collected in the DAVID tool knowledgebase (https://david.ncifcrf.gov/knowledgebase/DAVID_knowledgebase.html). The method identifies related genes by measuring the similarity of their global annotation profiles. So, the “grouping term” is based on the idea that two genes that have similar annotation profiles are functionally related. Each group term provides an enrichment score (ES) that indicates biological significance when ≥1.3 (equivalent to non-log scale 0.05). DAVID also provides a p-value to examine the significance of gene-term enrichment, which is corrected by Benjamini MT (Benjamini & Hochberg 1995).

Results

Discovery Phase

LASSO and ENET were applied to 20,899 gene expression probes in each of the three models. Under the conditions mentioned above, LASSO yielded 9 genes with a significant signal in the SNP models, 19 in the CpG models, and 23 in the Global models. In Table 3.2.1, we list the significant genes mapped to each probe with its deviance and p-value. Figures 3.2.2A, 3.2.2B, and 3.2.2C display all the probes analyzed with their deviances represented across the genome. Detailed information about the SNPs and/or CpGs mapped to these genes is provided as Supplementary Material (S3.2.1 – S3.2.6 Excel). ENET identified a lower number of significant genes: 11 in the SNP model, 6 in the CpG model, and 4 in the Global model. These results are shown in Table 3.2.2 and Figures 3.2.2D, 3.2.2E, and 3.2.2F. When the MT correction threshold was relaxed, ENET provided additional significant genes.

Some genes overlapped between methods and models: *CLIC6* was identified by the three LASSO models; *AIM2* and *SCNN1A* came out in the SNP and CpG models; four genes *PTN*, *CRTAC1*, *SERPINB3* and *SERPINB4* were identified in the SNP and Global models; and five genes (*S100A9*, *IGJ*, *FREM2*, *C15orf48* and *KRT20*) emerged in the CpG and Global models. Interestingly, 15 genes showed significance in the Global model when combining 3 *omics* data while they were not detected when analyzing only 2 types of *omics* data. The overlap of genes identified by the ENET model was lower: *MSMB* and *IGF2* were identified by the SNP and CpG models, and *PTN* and *SERPINB3* were selected by the SNP and the Global model. When comparing the methods we found overlap between LASSO and ENET in four (*PTN*, *SERPINB3, SERPINB4* and *CEACAM6*), one (*MSMB*), and three (*SERPINB3, PTN* and *IGHD*) significant genes in the SNP, CpG, and Global models, respectively. These results are displayed in Figure 3.2.3 using Venn diagrams. In the simulation study, as expected, no gene was significantly associated with any of the two methods and the three models. An example with LASSO and SNP model is shown in Figure S3.2.1.

Table 3.2.1. Statistically significant genes associated with SNPs and/or CpGs selected by LASSO&Permuted based maxT algorithm

Gene Name	Chromosome	Model	Deviance	p-value[1]
AIM2	1	SNPs	55.8	0.1
		CpGs	61.5	0.06
PLA2G2A	1	CpGs	71.4	0.01
S100A9	1	CpGs	53.7	0.03
		SNPs + CpGs	52.4	0.08
HMGCS2	1	CpGs	53.3	0.02
PIGR	1	CpGs	75.8	< 0.01
CTSE	1	CpGs	60.7	0.06
S100A2	1	SNPs + CpGs	58.7	0.04
CP	3	CpGs	51.1	0.02
TMEM45A	3	SNPs + CpGs	57.3	0.08
IGJ	4	CpGs	58.4	0.03
		SNPs + CpGs	59.0	0.09
UBD	6	SNPs + CpGs	75.0	0.07
TRIM31	6	SNPs + CpGs	47.1	0.1
PTN	7	SNPs	67.0	0.08
		SNPs + CpGs	92.0	< 0.01
ARHGEF35	7	SNPs + CpGs	49.6	0.09
CRH	8	SNPs + CpGs	56.7	0.1
CRTAC1	10	SNPs	66.2	0.03
MSMB	10	CpGs	67.3	0.06
CRTAC1	10	SNPs	60.9	0.1
		SNPs + CpGs		
TNNT3	11	CpGs	44.9	0.09
SAA1	11	SNPs + CpGs	127.8	0.04
SCCN1A	12	SNPs	57.9	0.08
		CpGs	58.8	0.03
KRT5	12	CpGs	58.2	0.03
TSPAN8	12	SNPs + CpGs	67.2	0.05
MYBPC1	12	SNPs + CpGs	74.5	0.08
SLC38A4	12	SNPs + CpGs	51.7	0.08
GTSF1	12	SNPs + CpGs	46.7	0.1

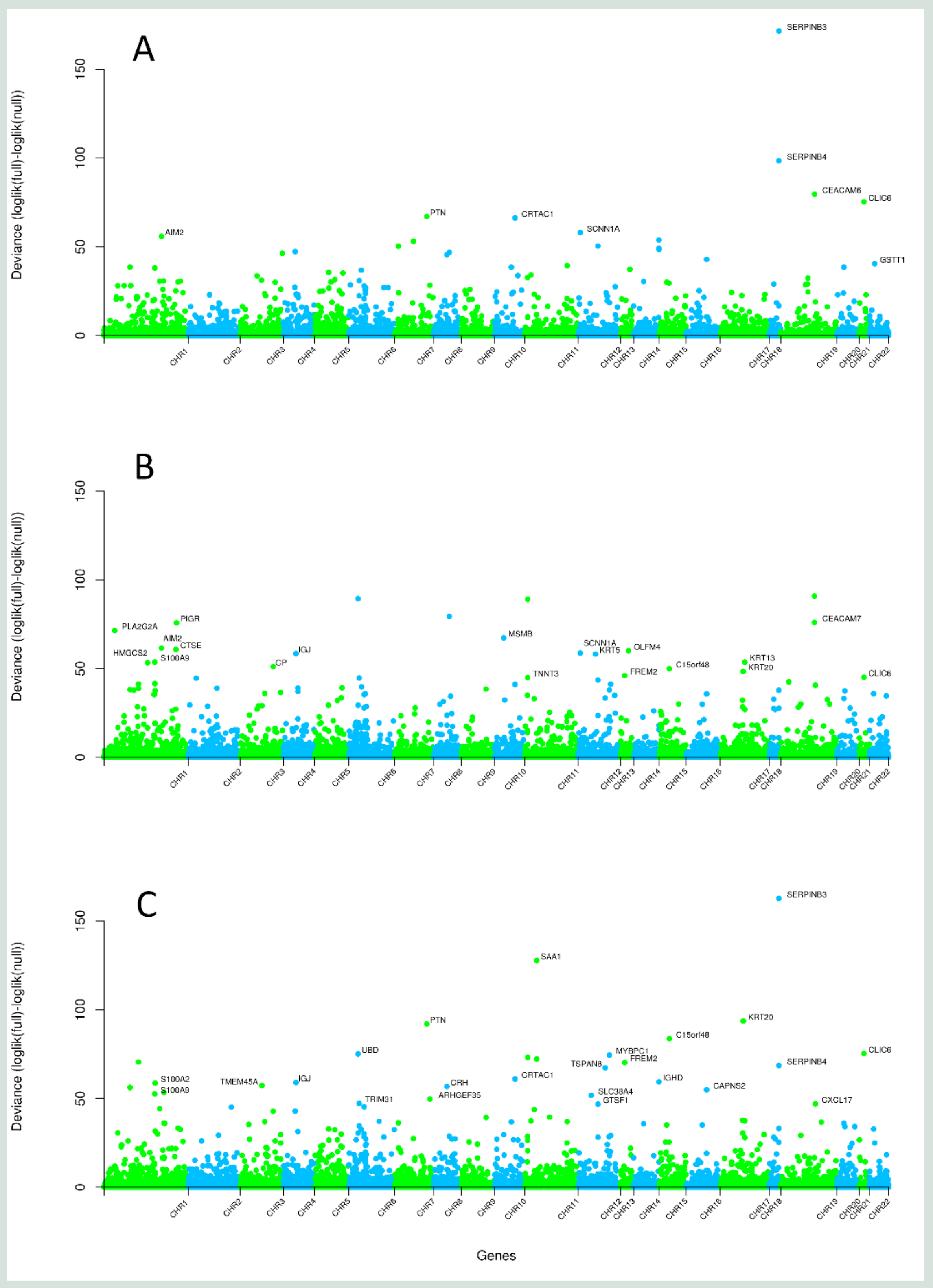
OLFM4	13	CpGs	60.0	0.06
FREM2	13	CpGs	46.0	0.06
		SNPs + CpGs	70.2	0.06
IGHD	14	SNPs + CpGs	59.4	0.1
C15orf48	15	CpGs	49.9	0.02
		SNPs + CpGs	83.7	0.05
CAPNS2	16	SNPs + CpGs	54.9	0.07
KRT20	17	CpGs	48.4	0.05
		SNPs + CpGs	93.7	< 0.01
KRT13	17	CpGs	53.6	0.02
SERPINB4	18	SNPs	98.4	< 0.01
		SNPs + CpGs	68.5	0.03
SERPINB3	18	SNPs	171.6	< 0.01
		SNPs + CpGs	162.7	< 0.01
CEACAM7	19	CpGs	76.0	< 0.01
CEACAM6	19	SNPs	79.6	0.01
CXCL17	19	SNPs + CpGs	46.8	0.1
CLIC6	21	SNPs	75.3	0.01
		CpGs	45.1	0.09
		SNPs + CpGs	75.3	0.07
GSTT1	22	SNPs	40.4	0.07

[1]The p-value was obtained after applying the permuted based – maxT algorithm and were therefore corrected for MT.

Table 3.2.2. Statistically significant genes associated with SNPs and/or CpGs selected by ENET&Permuted based maxT algorithm.

Gene Name	Chromosome	Model	Deviance	p-value[1]
REN	1	CPG	84.3	0.03
CRABP2	1	CPG	65.2	0.09
ANXA10	4	SNP	137.0	0.01
PTN	7	SNP	97.2	0.07
		SNP + CPG	102.5	0.09
MSMB	10	SNP	91.8	0.07
		CPG	78.9	0.06
MMP7	11	SNP	94.8	0.06
TCN1	11	SNP	88.9	0.07
IGF2	11	SNP	101.6	0.05
		CPG	92.1	0.04
MMP7	11	SNP + CPG	99.4	0.08
GTSF1	12	SNP	109.6	0.05
IGHD	14	SNP + CPG	97.5	0.1
SERPINB4	18	SNP	105.2	0.04
SERPINB3	18	SNP	171.6	0.02
		SNP + CPG	171.3	0.01
CEACAM6	19	SNP	108.4	0.03
NRLP2	19	CPG	84.2	0.04
CEACAM5	19	CPG	92.1	0.06
IGLJ3	22	SNP	97.7	0.05

[1] The p-value was obtained after applying the permuted based – maxT algorithm and corrected by MT.


A
Deviance (loglik(full)-loglik(null))
0
50
100
150
SERPINB3
SERPINB4
CEACAM6
CLIC6
GSTT1
PTN
CRTAC1
SCNN1A
AIM2
CHR1
CHR2
CHR3
CHR4
CHR5
CHR6
CHR7
CHR8
CHR9
CHR10
CHR11
CHR12
CHR13
CHR14
CHR15
CHR16
CHR17
CHR18
CHR19
CHR20
CHR21
CHR22
B
Deviance (loglik(full)-loglik(null))
0
50
100
150
PLA2G2A
PIGR
AIM2
CTSE
HMGCS2
S100A9
CP
IGJ
MSMB
TNNT3
SCNN1A
KRT5
OLFM4
FREM2
C15orf48
KRT13
KRT20
CEACAM7
CLIC6
CHR1
CHR2
CHR3
CHR4
CHR5
CHR6
CHR7
CHR8
CHR9
CHR10
CHR11
CHR12
CHR13
CHR14
CHR15
CHR16
CHR17
CHR18
CHR19
CHR20
CHR21
CHR22
C
Deviance (loglik(full)-loglik(null))
0
50
100
150
SERPINB3
SAA1
PTN
KRT20
C15orf48
UBD
MYBPC1
FREM2
TSPAN8
CLIC6
SERPINB4
S100A2
S100A9
TMEM45A
IGJ
CRTAC1
CRH
ARHGEF35
TRIM31
IGHD
SLC38A4
GTSF1
CAPNS2
CXCL17
CHR1
CHR2
CHR3
CHR4
CHR5
CHR6
CHR7
CHR8
CHR9
CHR10
CHR11
CHR12
CHR13
CHR14
CHR15
CHR16
CHR17
CHR18
CHR19
CHR20
CHR21
CHR22
Genes

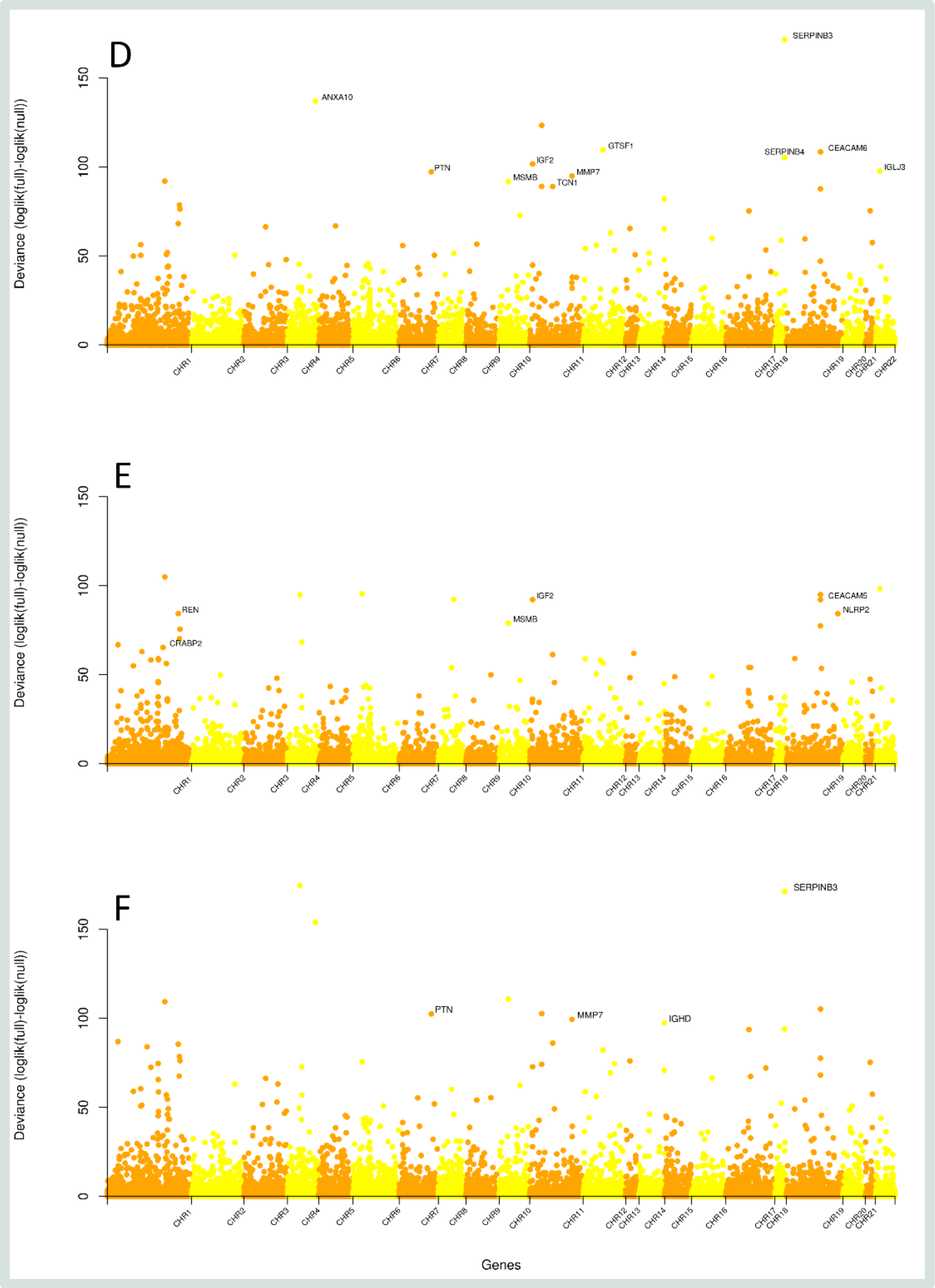


Figure 3.2.2. Deviance across the genome when applying LASSO and ENET to select SNPs, CpGs or both (Global model). The dots in the figure indicate the deviance of each gene located in the corresponding position in the genome. There are a total of 20,899 gene expression probes measured. Significant genes after applying the permutation-based MaxT method are tagged. The figures represent the deviance per gene expression probe using LASSO for the SNP model (A), the CpG model (B) and the Global model (C) and using ENET for the SNP model (D), the CpG model (E) and the Global model (F).

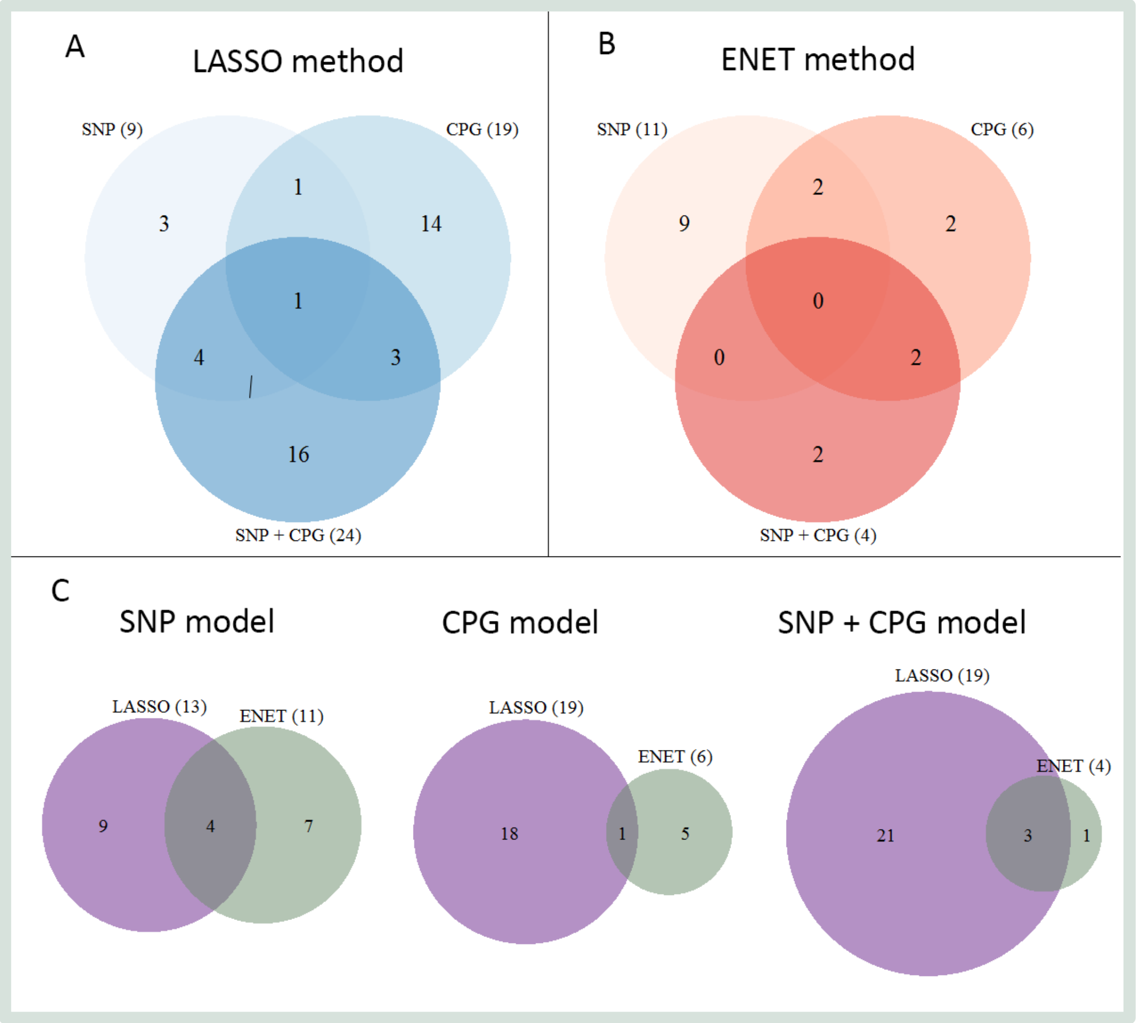


Figure 3.2.3. Venn diagrams showing the overlap between the significant genes compared by the two methods (LASSO and ENET) and models (SNPs, CpGs and Global). (A) Number of significant genes using the LASSO method for the three models (SNP, CPG, and Global); (B) number of significant genes using the ENET method for the three models (SNP, CPG and Global); and (C) number of significant genes per model comparing the two methods (LASSO and ENET).

Replication Phase

The replication study was restricted to those genes (n=48) that showed significant results in the discovery phase and we applied the same models, methods, and criteria of analysis to the TCGA data. Overall, we were able to replicate 75% of the results: 36 out of the 48 genes yielded a significant association at least in one of the models considered. Regarding the LASSO models, we replicated 3/9 genes from the SNP models, 17/19 genes from the CPG models, and 19/23 genes from the Global models (Table 3.2.3). Regarding ENET, we replicated 3/10 genes from the SNP model, 3/6 genes from the CPG model, and 3/3 genes from the Global model (Table 3.2.4).

Table 3.2.3. Significant genes obtained by LASSO&Permuted based maxT algorithm for the three models (SNP, CPG, and Global) in the original dataset (EPICURO Study) and the replication dataset (TCGA).

	Original Data (EPICURO)									Validation Data (TCGA)							
	Gene	probeset	Chr	Start	end	Dev	p-value	SNPs (N)	CpGs (N)	Dev	p-value[1]	SNPs (N)	CpGs (N)	SNPs (overlap)	SNPs (rep)	CpGs (overlap)	CpGs (rep)
SNP model	*SERPINB3*	8023696	18	61322433	61329197	171.6	<0.01	29		0	1	0		3	0		
	SERPINB4	8023688	18	61304495	61311502	98.4	<0.01	15		0	1	0		2	0		
	CEACAM6	8029098	19	42259398	42276113	79.6	0.01	10		0	1	0		0	0		
	CLIC6	8068383	21	36041688	36090519	75.3	0.01	30		3.5E-08	0.9	1		14	0		
	CRTAC1	**7935535**	**10**	**99624758**	**99790585**	**66.2**	**0.03**	**18**		**2.4E+09**	**0.001**	**12**		**4**	**1 (LD)**		
	GSTT1	**8074980**	**22**	**24376141**	**24384284**	**40.4**	**0.07**	**16**		**8.3E+07**	**<0.001**	**34**		**4**	**1 (LD)**		
	PTN	8143144	7	136912092	137028546	67.0	0.08	9		0	1	0		1	0		
	SCNN1A	7960529	12	6456011	6486523	57.9	0.08	26		0	1	0		8	0		
	AIM2	**7921434**	**1**	**159032275**	**159046647**	**55.8**	**0.1**	**6**		**5.7E+05**	**0.03**	**1**		**2**	**0**		
	CEACAM7	**8037053**	**19**	**42177235**	**42192096**	**76.0**	**< 0.01**		**19**	**1.5E+07**	**< 0.001**		**2**			**17**	**0**

CPG model	***PIGR***	**7923929**	**1**	**207101869**	**207119811**	**75.8**	**< 0.01**	**21**	**6.4E+08**	**0.001**	**19**	**18**	**1**
	PLA2G2A	**7913216**	**1**	**20301925**	**20306932**	**71.4**	**0.01**	**10**	**5.2E+09**	**< 0.001**	**57**	**9**	**0**
	CP	**8091385**	**3**	**148890292**	**148939832**	**51.1**	**0.02**	**3**	**1.6E+09**	**< 0.001**	**24**	**1**	**0**
	HMGCS2	7919055	1	120290620	120311555	53.3	0.02	8	0	1	**0**	8	-
	KRT5	**7963427**	**12**	**52908361**	**52914243**	**58.2**	**0.02**	**25**	**3.6E+12**	**< 0.001**	**112**	**24**	**5**
	C15orf48	**7983478**	**15**	**45722763**	**45725645**	**49.9**	**0.02**	**7**	**1.5E+08**	**< 0.001**	**23**	**5**	**0**
	KRT13	**8015323**	**17**	**39657233**	**39661865**	**53.6**	**0.02**	**8**	**8.2E+11**	**< 0.001**	**5**	**6**	**0**
	IGJ	**8100827**	**4**	**71521259**	**71532348**	**58.4**	**0.03**	**2**	**4.2E+08**	**< 0.001**	**19**	**2**	**0**
	SCNN1A	**7960529**	**12**	**6456011**	**6486523**	**58.8**	**0.03**	**29**	**2.1E+09**	**< 0.001**	**12**	**27**	**0**
	S100A9	**7905571**	**1**	**153330330**	**153333502**	**53.7**	**0.04**	**11**	**5.0E+11**	**< 0.001**	**33**	**9**	**1**
	KRT20	**8015124**	**17**	**39032141**	**39041495**	**48.4**	**0.05**	**3**	**5.9E+09**	**< 0.001**	**45**	**3**	**0**
	CTSE	**7909164**	**1**	**206317459**	**206332103**	**60.7**	**0.06**	**12**	**3.4E+09**	**< 0.001**	**36**	**12**	**1**
	AIM2	**7921434**	**1**	**159032275**	**159046647**	**61.5**	**0.06**	**8**	**4.7E+07**	**0.002**	**27**	**4**	**0**
	OLFM4	**7969288**	**13**	**53602972**	**53626186**	**60.0**	**0.06**	**10**	**1.6E+10**	**< 0.001**	**47**	**9**	**6**
	MSMB	7927529	10	51549553	51562590	67.3	0.06	7	0	1	**0**	6	0

	FREM2	**7968678**	**13**	**39261173**	**39461265**	**46.0**	**0.08**		**2**	**4.4E+07**	**< 0.001**		**13**			**1**	**0**
	CLIC6	**8068383**	**21**	**36041688**	**36090519**	**45.1**	**0.09**		**4**	**1.2E+08**	**< 0.001**		**19**			**4**	**0**
	TNNT3	**7937749**	**11**	**1940799**	**1959935**	**44.9**	**0.09**		**26**	**5.2E+08**	**< 0.001**		**72**			**22**	**0**
Global model	***SERPINB3***	**7920285**	**18**	**61322433**	**61329197**	**162.7**	**<0.01**	**15**	**0**	**3.0E+09**	**<0.001**	**6**	**4**	**1**	**0**	**0**	**0**
	KRT20	**7905571**	**17**	**39032141**	**39041495**	**93.7**	**<0.01**	**19**	**7**	**5.7E+09**	**<0.001**	**8**	**38**	**0**	**0**	**0**	**0**
	PTN	**7935535**	**7**	**136912092**	**137028546**	**92.0**	**<0.01**	**12**	**0**	**2.6E+08**	**<0.001**	**0**	**1**	**0**	**0**	**0**	**0**
	SERPINB4	**7938758**	**18**	**61304495**	**61311502**	**68.6**	**0.03**	**4**	**0**	**7.9E+08**	**<0.001**	**27**	**11**	**0**	**0**	**0**	**0**
	SAA1	7962559	11	18287808	18291521	127.8	0.04	20	1	7.9E+08	0.6	0	1	5	1	0	0
	S100A2	**7957966**	**1**	**153533587**	**153538306**	**58.7**	**0.04**	**20**	**7**	**1.0E+11**	**<0.001**	**1**	**5**	**5**	**0**	**3**	**0**
	C15orf48	**7964927**	**15**	**45722763**	**45725645**	**83.7**	**0.05**	**19**	**6**	**1.7E-07**	**<0.001**	**1**	**6**	**0**	**0**	**0**	**0**
	TSPAN8	7963817	12	71518877	71551779	67.2	0.05	8	1	9.9E+05	0.02	1	0	3	0	1	0
	FREM2	**7968678**	**13**	**39261173**	**39461265**	**70.2**	**0.06**	**14**	**2**	**2.9E+07**	**<0.001**	**3**	**10**	**3**	**0**	**1**	**0**
	CLIC6	**7983478**	**21**	**36041688**	**36090519**	**75.3**	**0.07**	**25**	**2**	**1.4E+08**	**<0.001**	**21**	**15**	**0**	**1 (LD)**	**0**	**0**
	UBD	**7995712**	**6**	**29523390**	**29527702**	**75.0**	**0.07**	**6**	**5**	**8.8E+08**	**<0.001**	**0**	**25**	**0**	**0**	**0**	**0**
	CAPNS2	**7981724**	**16**	**55600584**	**55601592**	**54.9**	**0.07**	**8**	**1**	**5.8E+07**	**<0.001**	**10**	**12**	**0**	**0**	**0**	**0**

MYBPC1	8023688	12	101988747	102079657	74.5	0.08	23	3	9.9E-08	1	0	1	2	0	2	0
TMEM45A	**8037197**	**3**	**100211463**	**100296285**	**57.3**	**0.08**	**12**	**0**	**1.6E+08**	**0.001**	**11**	**19**	**0**	**1 (LD)**	**0**	**0**
S100A9	**8015124**	**1**	**153330330**	**153333502**	**52.5**	**0.08**	**6**	**4**	**4.9E+11**	**<0.001**	**15**	**24**	**0**	**1 (LD)**	**4**	**1**
SLC38A4	**8023696**	**12**	**47158544**	**47219780**	**51.7**	**0.08**	**15**	**1**	**1.6E+08**	**0.001**	**8**	**15**	**6**	**0**	**1**	**0**
IGJ	8068383	4	71521259	71532348	59.0	0.09	3	2	3.3E+08	0.003	1	3	0	0	0	0
ARHGEF5	**8081288**	**7**	**143883177**	**143892791**	**49.6**	**0.09**	**8**	**0**	**1.2E+07**	**<0.001**	**11**	**8**	**0**	**1 (LD)**	**0**	**0**
CRTAC1	**8100827**	**10**	**99624758**	**99790585**	**60.9**	**0.1**	**7**	**5**	**3.8E+09**	**<0.001**	**7**	**9**	**1**	**0**	**3**	**1**
IGHD	8136981	14	106303102	106312014	59.4	0.1	7	1	-	-	-	-	-	-	-	-
CRH	**8151092**	**8**	**67088612**	**67090846**	**56.7**	**0.1**	**3**	**0**	**9.4E+08**	**<0.001**	**7**	**10**	**0**	**0**	**0**	**0**
TRIM31	**8178330**	**6**	**30070674**	**30080867**	**47.1**	**0.1**	**23**	**4**	**5.8E+08**	**<0.001**	**0**	**43**	**0**	**0**	**0**	**0**
CXCL17	**8143144**	**19**	**42932696**	**42947136**	**46.8**	**0.1**	**3**	**5**	**7.4E+08**	**<0.001**	**8**	**11**	**0**	**0**	**0**	**0**
GTSF1	**8124650**	**12**	**54849737**	**54867386**	**46.7**	**0.1**	**2**	**1**	**2.1E+07**	**<0.001**	**18**	**46**	**2**	**0**	**1**	**1**

[1] Bonferroni correction for the p-value were: 0.005 (SNP model), 0.003 (CPG model) and 0.002 (Global model); SNPs (N) and CpGs (N) are the number of SNPs and CpGs that were selected by LASSO per each gene expression probe in EPICURO data with the Illumina HumanHap 1M array and the Methylation 27k array; SNPs (overlap) and CpGs (overlap) are the number of SNPs and CpGs that were present in the TCGA data with the Genome wide 6.0 Affymetrix and the Methylation 450k array; and the SNPs (rep) and CpGs (rep) are the ones selected by LASSO in the TCGA data in common with the EPICURO data. The gene with "no p-value" is a gene that was not present in the RNASeqV2 in TCGA data.

Table 3.2.4. Significant genes obtained by ENET&Permuted based maxT algorithm for the three models (SNP, CPG, and Global) in the original dataset (EPICURO Study) and the replication dataset (TCGA).

	Original Data (EPICURO)									**Validation Data (TCGA)**							
	Gene	**probeset**	**Chr**	**Start**	**end**	**Dev**	**p-value**	**SNPs (N)**	**CpGs (N)**	**Dev**	**p-value[1]**	**SNPs (N)**	**CpGs (N)**	**SNPs (overlap)**	**SNPs (rep)**	**CpGs (overlap)**	**CpGs (rep)**
SNP model	**ANXA10**	**8098246**	**4**	**169013707**	**169108891**	**137.0**	**0.01**	**17**		**1.4E+08**	**<0.001**	**13**		**7**	**1**		
	SERPINB3	8023696	18	61322433	61329197	171.6	0.02	30		1.4E+09	0.08	32		3	0		
	CEACAM6	8029098	19	42259398	42276113	108.4	0.03	28		1.4E+09	0.04	4		5	0		
	SERPINB4	8023688	18	61304495	61311502	105.2	0.04	31		1.1E+08	0.07	10		8	1 (LD)		
	GTSF1	7963817	12	54849737	54867386	109.6	0.05	19		1.6E+06	0.08	7		9	2 (LD)		
	IGF2	**7937772**	**11**	**2150348**	**2170833**	**101.6**	**0.05**	**56**		**3.9E+12**	**0.002**	**31**		**12**	**0**		
	IGLJ3	7981730	22	23247030	23247205	97.7	0.05	183		-	-	-		-	-		
	MMP7	**7951217**	**11**	**102391240**	**102401478**	**94.8**	**0.06**	**19**		**2.8E+08**	**0.004**	**10**		**6**	**1 (LD)**		
	PTN	8143144	7	136912092	137028546	97.2	0.07	24		0	1	0		10	0		
	MSMB	7927529	10	51549553	51562590	91.8	0.07	78		0	1	0		0	0		
	TCN1	7948444	11	59620281	59634041	88.9	0.07	122		0	1	0		0	0		

CPG model	*REN*	7923608	1	204123944	204135465	84.3	0.03		22	0	1		1			22	0
	IGF2	**7937772**	**11**	**2150348**	**2170833**	**92.1**	**0.04**		**15**	**8.1E+12**	**<0.001**		**609**			**12**	**7**
	NLRP2	**8031398**	**19**	**55476652**	**55512508**	**84.2**	**0.04**		**34**	**8.0E+08**	**< 0.001**		**10**			**28**	**2**
	CEACAM5	8029086	19	42212530	42234436	92.1	0.06		26	9.3E+08	0.009		1			23	0
	MSMB	7927529	10	51549553	51562590	78.9	0.06		9	6.2E+07	0.3		36			7	1
	CRABP2	**7921099**	**1**	**156669410**	**156675375**	**65.2**	**0.09**		**39**	**1.1E+10**	**<0.001**		**132**			**35**	**11**
Global model	***SERPINB3***	**7920285**	**18**	**61322433**	**61329197**	**171.3**	**0.01**	**27**	**1**	**5.3E+09**	**<0.001**	**37**	**15**	**0**	**0**	**1**	**1**
	MMP7	**7951217**	**11**	**102391240**	**102401478**	**99.4**	**0.08**	**62**	**18**	**2.3E+08**	**0.003**	**5**	**2**	**0**	**0**	**0**	**0**
	PTN	**8143144**	**7**	**136912092**	**137028546**	**102.5**	**0.09**	**20**	**0**	**6.1E+08**	**<0.001**	**16**	**15**	**0**	**0**	**0**	**0**
	IGHD	7981724	14	106303102	106312014	97.5	0.1	35	6	-	-	-	-	-	-	-	-

[1] Bonferroni correction for the p-values is: 0.008 (CPG model) and 0.01 (Global model); SNPs (N) and CpGs (N) are the number of SNPs and CpGs that were selected by ENET per each gene expression probe in EPICURO data with the Illumina HumanHap 1M array and the Methylation 27k array; SNPs (overlap) and CpGs (overlap) are the number of SNPs and CpGs that were present in the TCGA data with the Genome wide 6.0 Affymetrix and the Methylation 450k array and the SNPs (rep) and CpGs (rep) are the ones selected by ENET in the TCGA data in common with the EPICURO data. The gene with no p-value is a gene that was not present in the RNASeqV2 in TCGA data.

Gene enrichment study

Using DAVID, 46 out of 48 genes showing significant signals in the discovery phase were annotated from 14 public categories. After enrichment analysis, 7 clusters with an ES ≥1.3 were found (S3 Table). The cluster with the highest ES (3.5) regarded to the terms "extracellular region, secreted, and signal peptide" grouping the genes *OLFM4, CRTAC1, MSMB, IGJ, MMP7, IGF2, PIGR, TCN1, CXCL17, S100A9, SAA1, IGHD, CRH, CTSE, FREM2, PLA2G2A, CEACAM7, CEACAM6, CEACAM5, REN, PTN, CP.* The rest of the clusters with an ES ≥1.3 were not significant after MT correction. Cluster 5 (ES=1.4) contains 3 genes coding for keratins (*KRT5, KRT13, KRT20*), cytoskeletal components that are regulated during urothelial differentiation, whose expression is altered in UBC, that have been proposed as markers for the molecular taxonomy of UBC (Choi et al. 2014). In addition, cluster 7 "EF hand and calcium ion binding" (ES= 1.3) contains multiple genes shown to play an important role in cancer (*S100A9, S100A2*, *CAPNS2, ANXA10, CRTAC1, FREM2, MMP7, PLA2G2A*), including two members of the S100A family of proteins.

Discussion

Integration analysis is an emerging area in the field of *omics* data analysis to find new biological insights into complex traits (Knowles & Hurst 2014). In this regard, our pathophysiological understanding of cancer could be improved by using innovative approaches based on *omics* data to identify hidden mechanisms in which multiple factors are involved. We previously analyzed the set of *omics* data used here following a multi-stage approach by proposing an *omics* integration analysis framework. The results of this previous work highlighted relevant *omics* trans-acting relationships in UBC (Pineda, et al. 2015). Here, we propose an *omics* integrative analysis pipeline using LASSO and ENET, and focus on cis-acting relationships that appear to have a predominant role in the regulation of gene expression (Leung et al. 2015). The three *omics* data are combined in a large input matrix and then a permutation-based MaxT method is adapted to assess the significant models while correcting for MT.

In comparison with classical approaches (Kristensen et al. 2014; Ritchie et al. 2015), our strategy has several advantages, including the possibility of working with a large number of parameters, even if the sample size is small, dealing with more than one set of heterogeneous data with highly-correlated variables, and providing MT corrected p-values to assess the models' goodness of fit. Furthermore, the results are easily interpretable due to the dimensionality reduction during the variable selection process.

The expression of 48 genes was found to be significantly associated with SNPs and CpGs in UBC, pointing to new mechanisms in an intricate scenario where common genetic variants and DNA methylation regulate gene expression in cis-acting (1MB) relationships. Some of the genes were identified by the three models and by the two methods, likely underscoring the existence of true relationships.

The application of LASSO and ENET as part of the aforementioned integrative analysis framework led to different results. This is not surprising, mainly for two reasons: (1) the α parameter (equation 5) used by LASSO is always equal to 1 while ENET uses $\alpha < 1$. This gives a smaller penalization and therefore more variables with $\beta \neq 0$ were foreseen using ENET; and (2) the fact that SNPs and CpGs may be correlated, mainly when they are closely positioned in the genome, leads LASSO to select one from the set of parameters that are highly correlated while ENET forms groups of nets with these variables. In our analysis, only 4/24 (SNP model), 1/25 (CpG model) and 3/28 (Global model) genes were shared by both methods. The genes detected only by LASSO showed large deviances and borderline p-values with ENET. Waldmann *et al* (Waldmann et al. 2013) reported that ENET usually detects more true and false positive associations. In our case, this may result in an increased probability of having significant associations by chance. In turn, this can lead to reduced power. On the other hand, ENET selected some genes that were not selected by LASSO, mainly due to the correlated structure of the parameters. An example of this is displayed in Figure 3.2.4, showing that *MMP7* has three correlation nets that probably are responsible for the gene selection with ENET and not with LASSO. These comparisons are shown in Table S3.2.2.

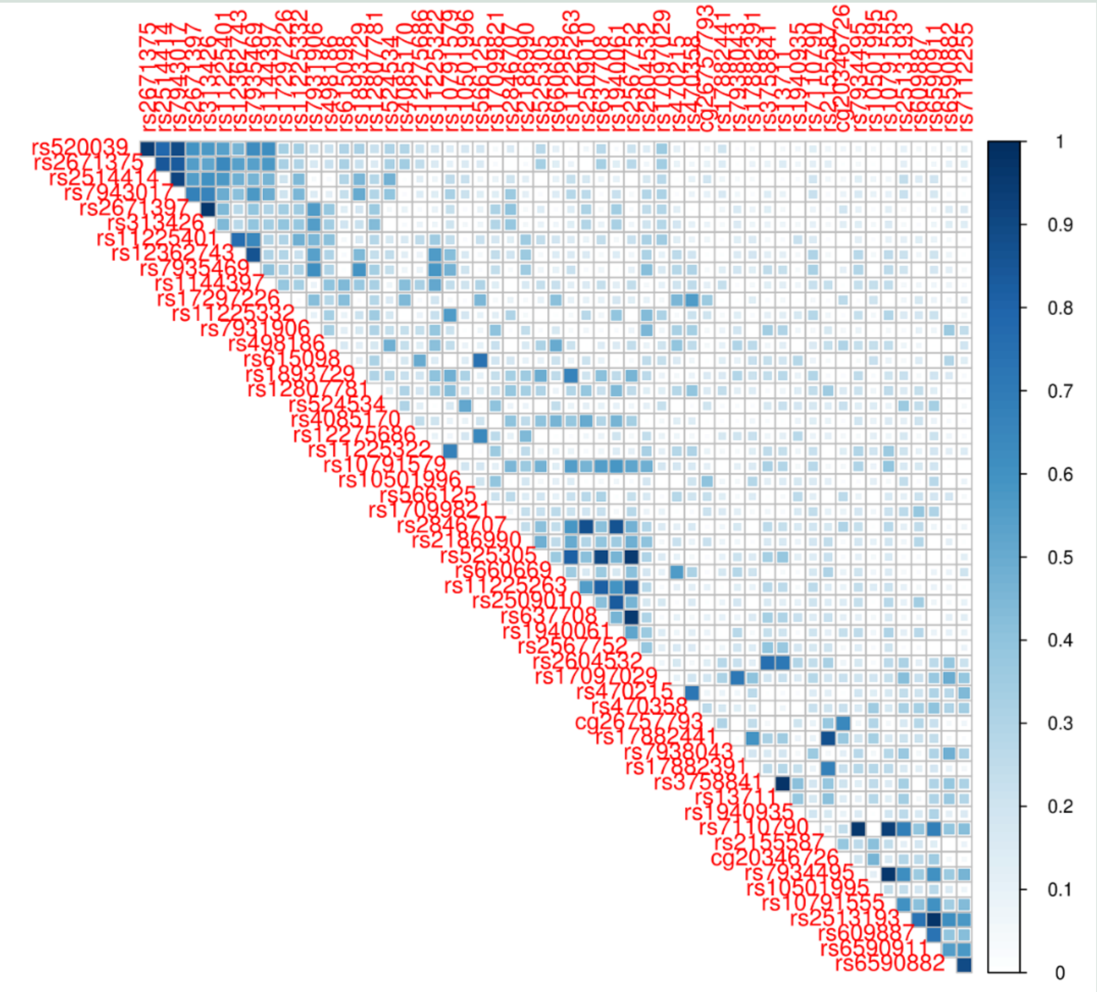


Figure 3.2.4. Example of a correlation plot for *MMP7* detected by the Global model using ENET but not using LASSO. The bar color represented the levels of correlation from 0 (no correlation) to 1 (perfect correlation) between SNPs and CpGs that were selected for the *MMP7* models. Three nets of correlated variables are the ones responsible that the gene is only selected by ENET and not by LASSO.

Regarding the differences between the models, 13/25 and 6/20 significant genes in the CpG and 6/20 the SNP models, respectively, were not significant in the Global model. It is reported in the literature that 10% of SNPs are associated with gene expression and DNA methylation (Wagner et al. 2014; Bell et al. 2011), hence DNA methylation may confound or modify the association between SNP and gene expression. Even though this is a potential explanation, discordances resulting from sample size cannot be discarded since the penalty function is selected by CV. However, k = 5 was used to apply the k-fold CV to decrease the problem of small sample size without increasing bias. In the reverse scenario, 16 genes were selected exclusively in the Global model. Some of the genes identified had high deviances and borderline p-values, probably because the Global models increase the deviance due to the addition of more information when integrating data. For the non-significant genes, the explanation could be the existence of an interaction effect between SNPs and CpGs (Table S3.2.3). This further supports the importance of integrating *omics* data to discover hidden information.

The validity of the strategy that we have developed, and of the results obtained, is supported by the fact that 75% (36/48) of the genes identified in the discovery phase were replicated using TCGA data by applying the same strategy. This represents 64% of all gene models found since some of the genes overlap between models and approaches. Also, the null results of the simulation study indicate that the significant associations found are unlikely to be due to chance.

Importantly, several of the genes that emerged from our analyses have been previously shown to be important in bladder cancer biology, including *KRT20, IGF2, CTSE, ANXA10* and *CRH.* These genes have already been proposed for a panel of molecular markers to improve the diagnosis and follow-up of UBC as part of a 12-gene expression urine signature to identify patients suffering from UBC and predict tumor aggressiveness (Mengual et al. 2010). The five genes aforementioned were also replicated in the TCGA data. Furthermore, *KRT20, IGF2,* and *CTSE* have also been previously associated with UBC. *KRT20* is a highly specific marker of umbrella cells in normal urothelium and its expression is commonly altered in papillary non muscle-invasive UBC as well as in muscle-invasive UBC. It has been proposed that the correlation between *FGFR3* mutations with normal *KRT20* expression pattern may indicate that the mutation occurs earlier (van Oers et al. 2007). Loss of imprinting (LOI) is a common epigenetic event in cancer and a LOI of *IGF2* has been reported in UBC (Byun et al. 2007). In our analysis, *IGF2* was detected in the SNP and CPG models suggesting that both type of factors may be involved in regulating the expression levels of this gene. *CTSE* expression was significantly associated with progression-free survival in pTa tumors in a study of gene expression profiles in UBC (Wild et al. 2005).

We also performed a gene enrichment analysis to assess whether the significant genes had related biological functions. The cluster with the highest ES was ”Extracellular region, secreted, and signal peptide”. Secreted proteins are known to play a crucial role in cell signaling and the cellular secretome has a major impact on multiple aspects of tumor cell biology (cell growth, migration, invasion, and angiogenesis) (Karagiannis et al. 2010). One cluster highly enriched in keratins points to the regulation of cell differentiation, known to be important in the molecular classification of UBC. In addition, some genes - including *S100A9* and *S100A2* - were grouped under the "EF hand and calcium ion binding" term. The *S100* family is composed of, at least, 24 members carrying the Ca^{2+} binding EF-hand motif. Expression of *S100* protein family members is regulated during inflammation and carcinogenesis and has been associated with poor prognosis in patients with UBC (Yao et al. 2007). Other studies have reported an overexpression of *S100A9* in UBC tissue (Dokun et al. 2008; Minami et al. 2010).

Limitations of this work are the small sample size of the discovery phase study, due to the lack of enough fresh tumor tissue from the same set of individuals, and the lack of a comparable and independent UBC patient series with the 3-*omics* data available to replicate our results. While the discovery EPICURO study recruited all patients with UBC, the TCGA project focused on muscle-invasive UBC. In addition, different high-throughput technologies/platforms were used in each of the studies. The SNP arrays genotyped different SNPs and, consequently, provided different genomic coverage. The TCGA used a DNA methylation array of 450k with much higher resolution than the 27k the one used in the EPICURO study. Finally, the use of different technologies to measure transcriptomics is a considerable limitation. In the EPICURO discovery phase, gene expression levels were measured with microarrays which provide relative values at probe set level, that is, for one gene different expression levels can be obtained from each mapped probe, while in the TCGA study gene expression was measured with RNA-seq which gives absolute gene expression values. These differences between data sets introduce a massive heterogeneity that makes the replication even more difficult. In spite of that, we replicated 75% of the identified genes (64% of the models) with TCGA data, providing strong support to the appropriateness of our approach and the relevance of the results obtained. Another potential limitation is the fact that tumor samples are heterogeneous regarding neoplastic cell content and stromal cell composition. Consequently, we checked the expression of all significant genes in a panel of UBC cell lines with available microarray expression data (Earl et al. 2015) and found that all but one (*IGJ*) are expressed in urothelial tumor cells, indicating that our analyses likely reflect genomic regulatory events in the tumor cells. It is, however, likely that relevant genomic interactions control gene expression not only in neoplastic cells but also in the stroma. Given the importance of the latter in tumor progression, further integrative *omics* studies using microdissected material will be highly informative.

One important strength of the approach used here is the lack of need to filter by LD in SNPs, or grouping CpGs within CpG islands, when dealing with a huge number of heterogeneous and correlated parameters delivered by different arrays. This emanates from the fact that LASSO and ENET can deal with highly correlated variables while performing variable selection. By performing data reduction/filtering before applying the statistical methods, there is a chance to filter out the functional SNPs and/or CpGs and thereby lose their association with gene expression. The adaptation of a strategy that performs a permutation and the maxT algorithm to assess p-values and to correct by MT, avoiding a double permutation and therefore reducing computational time, is also worthwhile emphasizing. In this regard, the permutation-based method considers the permutation of individuals within each gene, allowing to control for the possible dependence structure between genes. In addition, the MaxT algorithm is a permutation-based FWER controlling procedure which is adapted to the correlation structure found in the data and has been shown to be asymptotically optimal under dependence (Meinshausen et al. 2011).

In summary, we demonstrate that the integration of multiple *omics* data types allows the identification of hidden mechanisms that were missed when analyzing single *omics* data types individually. There is an urgent need to develop statistical methods to fill the gap between the huge amount of data generated and the mechanistic understanding of complex diseases. Here, we present two penalized regression methods (LASSO and ENET) in combination with a permutation – based strategy (permutation-based MaxT method) to deal with common problems found in integrative analysis: heterogeneity between data types, number of individuals much smaller than the parameters to assess, multicollinearity, and sparseness to facilitate the interpretation of the results. This approach is flexible and easy to implement in different *omics* data and diseases as well as when considering interaction terms in the model. We contribute to the field with a methodological development and with several significant and sound molecular associations conforming part of the genetic architecture of UBC. By using this cancer as an example, we conclude that modeling the intricacy of *omics* data variation with appropriate statistical strategies will certainly improve our knowledge of the mechanisms involved in complex diseases.

Responses to the Reviewer's comments

Reviewer #1

Reviewer #1: Pineda et al use Lasso and elastic net with correct for multiple testing with MaxT to integrate several genomics platforms. 27 patients from the Spanish Bladder Cancer EPICURO study had SNP data, DNA methylation and gene expression data for testing the method. This is an interesting approach that could also be used to model subtypes and outcomes.

(1) For lasso and elastic net, you can adjust the parameters of alpha and lambda. Is there a big difference in the parameters between the individual SNP and individual CpG analysis vs the combined analysis?

It would be good to add to the supplement the parameters used. Is your elastic net model leaning towards lasso or ridge regression?

For each model assessed, the parameters were re-estimated using cross validation. To obtain the lambda by LASSO penalty, a 5-fold cross validation was applied maximizing the penalized log-likelihood function. For ENET, the optimal penalty for lambda was obtained using the same strategy (5-fold cross validation maximizing the penalized log-likelihood) but in this case we selected the best alpha using an alpha vector of α ϵ (0.01, 0.99) by 0.01 as explained in Material and Methods, line 130 (LASSO) and line 139 (ENET). So, we obtained one estimation per model and method making a total of 125,394 lambdas and 62,697 alphas. Therefore, it was not possible to study one by one but checking globally if there were patterns of the parameters lambda and alpha that deserve further attention.
Regarding the differences observed between models, for LASSO the mean(lambda) = 0.135, 0.08 and 0.129 for SNP, CPG and Global model respectively. For ENET, the mean(alpha) = 0.46 and sd(alpha) = 0.31 with minimum value = 0.01 and maximum value = 0.99 for all of the models (SNP, CPG and Global model). Therefore we could not assume any pattern or leaning towards lasso or ridge regression between the three models and approaches.

If the tuning parameters for SNP were drastically different from what was seen for CpG, could you bias feature selection in the combined model if a suboptimal alpha and lambda was picked for one or both of the data platforms. In tables 3 and 4 it looks like there is a preferential pick of SNPs over CpGs in the combined model.

The tuning parameters between SNP and CPG models were different because this is the way that penalized regression methods control for different scales. SNPs are categorical and CpGs are continuous in our analysis, thus it is normal to have different tuning parameters. In our previous answer we showed that Global model had a tuning parameter in the middle but closer to the SNP model at global numbers. The tuning parameter is not the responsible for selecting more SNPs but it is closer to the SNP model because more SNPs are selected. In any case, the numbers showed are global and we cannot extract any conclusion without studying individually one by one, a task that becomes impossible because of the amount of lambdas and alphas estimated. This question of big numbers is one of the challenges we try to answer in this paper when analyzing high-throughput data.

A potential explanation to the question of why the Global model selected more SNPs than CpGs is stated in the manuscript (Page 31). In the introduction we hypothesized that the biological idea behind the integration is based in previous finding where gene expression was affected by both DNA methylation and genetic variants, both co-regulating different genome spaces interpreting that the regulation of expression of a given gene results from the combination of genetic variants that, at the same time, could be influenced by the levels of DNA methylation in specific CpGs. In Results we showed that, when integrating, different associations are found and in the Discussion we commented on the differences and its potential causes such as correlation, confusion and/or modifier factors. We do not detail each specific example to not enlarge the length of the manuscript but, in supplementary material, we provide the entire list of genes, SNPs, and CpGs per model and method to allow the readers go through these details if interested.

(2) I highly recommend a supplemental table that lists the samples used for analysis from both the EPICURO and the TCGA study so that people could have the exact data set if they want to implement and test your method. It would be useful since your performed your analysis on 27 of the 70 patients in EPICURO and 238 of the over 400 TCGA samples.

Following the reviewer's recommendation we have added S1 Table with information on the 27 samples from EPICURO data and S2 Table with information on the 238 samples from the TCGA data used in the analysis. With this change, the previous S1-S3 Tables are now S3-S5 Tables.

(3) Gene expression was modeled by your methods with SNP data alone, CpG data alone, and SNP and CpG data combined. Did you limit to the same 27 EPICURO patients for all analyses?

Yes, we used the same population in the three models to make them comparable. S1 and S2 Tables specify now the samples we have used.

(4) The Discovery phase was performed on a probe level for the gene expression data vs the replication phase which was performed on a gene-level. In your final 48 genes selected from your model, where there probes for the same gene (from the microarray) that were discordant? If there were, could these be some that didn't replicate in the TCGA data.

It is very well pointed that the discordance in the replication phase could be explained by differences between the approaches used to measure gene expression levels in the two phases. We comment on this in the Discussion, line 480, one of the limitations of this type of analysis being the considerable heterogeneity introduced by the different methods applied by the replication study.
As the reviewer points correctly, we observed that from the set of the 48 genes significantly associated in the EPICURO data (microarray technology), four were mapped for more than one probe, from which only one gene (*SAA1*) was not replicated in the TCGA data (RNAseq). One possible reason could be that microarray technology maps the probes to the genes for different transcripts given relative values of gene expression while RNAseq gives the absolute value of gene expression. Consequently the expression levels observed by one technology cannot be observed by the other. Effectively, *SAA1* had a Pearson coefficient = 0.58 between probes while the other three had Pearson coefficients = 0.94, 0.90 and 0.98. This illustrates one type of the problems found in replicating when data heterogeneity is introduced. We have now specified in the Discussion, line 479, the differences between microarray and RNAseq.

(5) There were a few genes that made the SNP only and the CpG only models but not the combined model. Any thoughts as to why?

In line 421, we discussed that, in the literature, 10% of SNPs are associated with gene expression and DNA methylation, so DNA methylation may confound or modify the association between SNP and gene expression. Although, we believe this is one possible explanation, we cannot discard small sample size may also be an issue since we estimated the optimal parameters using k-fold cross-validation. In any case we used the smallest k possible (k=5) to apply cross validation to avoid sample size problems in variable selection while not introducing big bias. We have clarified the k used in line 143 and further comment on sample size limitation when using cross validation in the Discussion, line 427.

(6) It would also be useful to know if these are muscle invasive or non-muscle invasive or histology designations such as papillary and squamous. The fact that only 75% of your genes were replicated in TCGA could be because of a different patient population in addition to the difference in data platforms.

In line 473, we discuss about this issue since the TCGA samples are only muscle invasive UBC while in EPICURO are both, muscle-invasive and non-muscle invasive. Consequently this could also be one of the reasons why the replication is only 75% of the genes. To clarify these differences we have now added in Material and Methods, lines 170 and 212, the information from both EPICURO data and TCGA data.

Reviewer #2

Reviewer #2: In this manuscript, the authors describe a permutation-based algorithm for assessing the significance of relationships uncovered by penalized regression methods (LASSO and ENET) in “multi-omic” databases, concomitantly using the MaxT algorithm to correct for multiple hypothesis testing. The method is applied to combined mutation, DNA copy number, and gene expression data from the authors’ own urothelial bladder cancer (UBC) studies and “replicated” by analysis of the TCGA UBC data. They then follow up with simulation studies and a Gene Ontology enrichment analysis. As stated in the cover letter, the principal contribution is methodological. It follows up on a more biologically-focused version of the analysis by the authors (Pineda, et al., Human Heredity, in press).
None of the individual ingredients of the method are novel, but they are combined (and customized for the particular types of omic data) in an interesting way. Penalized regression methods should probably be used in omic analysis even more often than they are -- because of the typical mismatch between number of variables and number of cases. And careful attention to multiple hypothesis testing is vital to proper control of Type I statistical error.

(1) I don't THINK that lack of independence of the vectors over genes or unequal variances are issues that would compromise the validity of the permutation test in this algorithm. But, since the authors have presumably thought deeply about those questions, they should address them briefly but directly -- at a place of their choosing in the manuscript.

The permutation test was done permuting individuals within the genes with the intention of avoiding any problem of dependence structure between genes. When we considered to apply a permutation-based approach to assess significance, we thought in two possibilities, one permuting individuals within each gene and two, permuting genes within each individual. The results from the second approach were more restrictive due to the potential correlation structure of the genes since gene expressions may be correlated between them. For that reason, we decided to use the permutation-based approach permuting individuals. In addition to this, the MaxT algorithm is a permutation-based FWER controlling procedure that is adapted to the correlation structure found in the data and has been shown to be asymptotically optimal under dependence. Following the reviewer's recommendation we briefly and directly comment on this in the Discussion, line 514, by adding also a new reference (Meinshausen et al. 2011)

(2) The level of writing is uneven. I suggest careful copy-editing of the text prior to publication.

We have carefully edited the manuscript.

(3) Relatively minor comments/questions:

1. Line 244: I assume that the gene enrichment analysis using David was based on the Gene Ontology. If so, that should be stated directly, along with specification of the database type and evidence parameters used for the analysis.

The gene enrichment analysis is based on 14 annotation categories (Gene Ontology (GO), Biological process, GO Molecular Function, GO Cellular Component, KEGG Pathways, BioCarta Pathways, Swiss-Prot Keywords, BBID Pathways, SMART Domains, NIH Genetics Association DB, UniProt Sequence Features, COG/KOG Ontology, NCBI OMIM, InterPro Domains, and PIR Super-Family Names) collected in the DAVID knowledgebase (https://david.ncifcrf.gov/knowledgebase/DAVID_knowledgebase.html).
Even though this information can be found in the reference we provide in the manuscript, we specify now in the manuscript line 267.

2. Line 312 – 320. "Importantly, we replicated results for 36 (75%) of them in an independent data set (TCGA)." Does "restricting the analysis to those genes …" mean that the calculation was done over the entire set but that only the results for the 48 genes (i.e., 75% replication) are being reported here? The figures given would seem to relate to the sensitivity of replication but not its specificity. What about the numbers related to specificity? Overall, I think the replication study needs further description and (no pun intended) specificity.

The replication phase is done by restricting it to all significant genes detected in the discovery phase (48 genes) considering all models (SNP, CPG and/or Global) and methods (LASSO and/or ENET). That is, we focused only to the genes significantly associated in the discovery phase and applied the same strategy used in the first phase to check whether they were also significantly associated with the SNPs and CpGs in the TCGA data. Following the reviewer's suggestion, we have now further commented this point in Material and Methods, line 250, to better clarify about the replication effort.

Regarding sensitivity and specificity, we do not consider the 75% of replicated genes as a sensitivity estimation rate because we cannot consider our results as a "gold standard" and therefore, the estimation of specificity is behind the same consideration. In spite of this, to be sure we were doing the replication properly, we performed the same replication study with a set of genes randomly selected from the TCGA data. We observed that the percentages of significant genes were much lower than when considering the significant genes in our discovery phase. However, sample size is much higher in TCGA, so we cannot differentiate between a true positive due to power issues, a true positive due to a threshold selection, and a false positive in the replication dataset.

It's appropriate that the authors list the major differences in the two studies (platform, etc.) and that the level of concordance is perhaps surprising.

We were conscious of the different platforms used in the two studies, among other methodological aspects, was a limitation and we believe that this is one of the reasons that may explain why we do not replicate the 100% of the associations at the gene level. We already commented about this limitation in line 474. However, as suggested by the reviewer, we have now added in this part of the Discussion more details about the different platforms used in both studies to further point out this limitation. Regarding the level of concordance, it should be consider that the small sample size in the discovery phase may also be an important limitation and those significant genes identified are probably the ones with the highest signal.

3. Lines 337 – 346: This manuscript is focused on methodology, not biology, but, nonetheless, the reader will wonder whether there's any biological significance to the categories that showed up in the gene ontology analysis. Do the authors think there's any meaning to the categories or are they just ones that happened to show up despite the multiple hypothesis-testing correction?

Indeed, the genes identified correspond to pathways or processes that are known to be important in bladder cancer. We did not wish to emphasize this too much in our previous version of the ms. but we appreciate that this comment gives us the opportunity to make this point further and we have done so in the last section of Results as well as in the Discussion.

4. Lines 371 – 381: Could another reason for the lack of concordance between LASSO and ENET be the sheer statistical arbitrariness of selecting just a handful of top genes out of many thousands -- especially given the occurrence of high correlations among the vectors for different genes?

The necessity of giving a threshold to correct my multiple testing is always problematic. The equilibrium between type I error and type II error is one of the most important issues in statistical genetics. Which is the threshold of false positives we are able to assume while no detecting false negatives? One of the differences we observed that can be seen in Table S4 (Table S2 before) and we comment in the Discussion, line 405, is that some of the genes selected by LASSO are borderline with ENET. So the arbitrariness of the statistical threshold is an important issue. In the reference we provide, Waldmann et al. demonstrated that ENET usually detects more true and false positives associations and can provoke a decrease in power.

5. Line 449. What does it mean in terms of the data to say that only one gene was "found not expressed in UBC cell lines". What cell lines? Below detection limit? How does "being expressed" relate to enrichment?

Chapter 3. Integrative eQTL –omics analysis considering tumor tissue and blood samples in individuals with bladder cancer

Silvia Pineda (1,2), Kristel Van Steen* (2,3), Núria Malats* (1),

(1) Genetic and Molecular Epidemiology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.

(2) Systems and Modeling Unit – BIO3, Montefiore Institute, Liège, Belgium.

(3) Systems Biology and Chemical Biology, GIGA-R, Liège Belgium.

* Equal contributions

American Journal of Human Genetics 2015 – Submitted

Abstract

Integrative –omics analysis approaches to combine different data are emerging. These data are mainly determined in the same source material (i.e., tumor sample). However, the integration of data from different material levels (i.e., blood and tumor) may also reveal important knowledge on the human genetic architecture. To model this multi material-level structure, integrative-eQTL analysis applying 2-Stage Regression (2SR) has been proposed. This approach consists on two stages, (1) gene expression levels are regressed with markers at somatic level and (2), the residuals-adjusted are regressed with the germline genotypes. Such an approach relies on several assumptions needed to overcome challenges high-throughput -omics data impose. Previously, we have shown that penalized regression methods in combination with a permutation–based MaxT method are promising to this regard. In this report, we assess whether our previously developed strategy can also be considered when integrating different data source material and we compared it with two different ways of parameter estimation in the 2SR-approach, one using multiple linear regression (MLR) and other using LASSO to control for correlated data. We applied the three strategies to integrate genomic, epigenomic and transcriptomic data from tumor tissue with germline genotypes from 181 individuals with bladder cancer from the TCGA Consortium. Our study showed no significant results when the 2SR-MLR was applied supporting, as previously showed, the underestimation of this approach when variables are correlated, in contrast of the other two approaches. Furthermore, our approach propose a list of relevant eQTLs including and extending the ones found by the 2SR-LASSO approach to be considered in future analysis.

Integrative –omics data analysis (IODA) are starting to emerge for the combination of different high-throughput platform biological derived information to better understand the complexity of biological systems. IODA are mainly focused in the combination of –omics data from the same source material, such as tumor samples (Mankoo et al. 2011), human brain (Zhang et al. 2010; Gibbs et al. 2010) or blood samples (Van Eijk et al. 2012). However, the integration of data from different material levels (i.e., blood and tumor) may also reveal important knowledge about the hierarchy of the human genetic architecture. For instance, the influences of germline variants in gene expression are normally studied in cell lines and normal tissue but not in tumor tissue due to the complexity of the transcript regulation caused by the somatic changes produced in the tumor (Fredriksson et al. 2014). To approach this further level of complexity, Li *et al.* (Li et al. 2013; Li et al. 2014) performed *cis*-eQTL analysis in breast cancer by applying a 2-stage regression (2SR) approach. In stage one, the gene expression level was regressed with all somatic changes (copy number variation and DNA methylation) in *cis*-relationship. In stage two, the residual-adjusted outcome was regressed with the germline genotypes also in *cis*-relationship. While this approach presents computational and data management advantages, its validity relies on assumptions that are not met in the majority of large scale genomic studies, namely the risk factor of interest should not to be correlated with the rest of the covariates and the tumor SNPs and/or CpGs may be highly correlated between them. When this occurs, the 2SR approach may introduce a bias resulting in a loss of power as demonstrated by Demissie *et al.*(Demissie & Cupples 2011) and Che *et al.*(Che et al. 2012) in comparison with a multiple linear regression (MLR) approach. Similarly, an application of the 2SR approach for detection of QTLs has shown a loss of power when covariates are correlated (Zeegers et al. 2004). Another limitation of the method proposed by Li *et al*. is that they applied MLR, a method that cannot be applied when the number of parameters to be estimated in the model is larger than the sample size (n<<p problem), a prevalent scenario in large genomics studies. Demissie *et al*(Demissie & Cupples 2011) and Che *et al.*(Che et al. 2012) also mentioned on the critical issue when applying 2SR and MLR because when a study involves correlated independent variables, the two approaches produced incongruent results. This is a fact in the majority of the genomic analyses, since variables such as SNPs and CpGs may be highly correlated between them and therefore the eQTLs may suffer of harmful multicollinearity, thus, none of the two approaches are valid.

To overcome these limitations, we propose to adapt an integrative method we previously developed and which is based on penalized regression (LASSO and ENET) in combination with permutation-based maxT method(Peter H. Westfall & Young 1993) to obtain p-values and correct them for multiple testing (Pineda, et al. 2015). This approach can deal with the challenges that large genomic studies impose and it is well suited to perform eQTL assessment when –omics multi material-level data need to be integrated. Here, we apply our method by performing an eQTL- IODA analysis and compare our strategy with the 2SR approach and with a 2SR using LASSO to deal with the n<<p problem and correlated variables.

Urothelial bladder cancer (UBC) tumor data and blood sample from the same patients were obtained from The Cancer Genome Atlas (TCGA) consortium (https://tcga-data.nci.nih.gov/tcga/) by downloading and processing them with the TCGA-Assembler(Zhu et al. 2014). The data was profiled with Genome wide 6.0 Affymetrix, RNASeqV2, and the HumanMethylation450K Illumina array. A total number of 905,422 SNPs, 20,502 gene expression probes, and 350,271 CpGs were obtaining from tumor samples and 905,422 SNPs from blood samples for 181 individuals. SNPs were measured with the same platform in blood and tumor samples expecting to have some differences due to somatic mutations within the tumor. Consequently, we performed an agreement study by pairs of SNPs to check the rate of disagreement between tumor and blood using the weighted Kappa Index (wKI). Each pair was represented in a weighted matrix where cells located on the diagonal represent agreement, while cells one off the diagonal are weighted 1, and cells two of the diagonal are weighted 2. Figure 3.3.1 displays the percentage of disagreement (wKI) per chromosome. The highest disagreement is observed in chromosome 9 (29%) supporting the deletion of both arms of the chromosome 9 frequent in bladder cancer patients (Wu 2005).

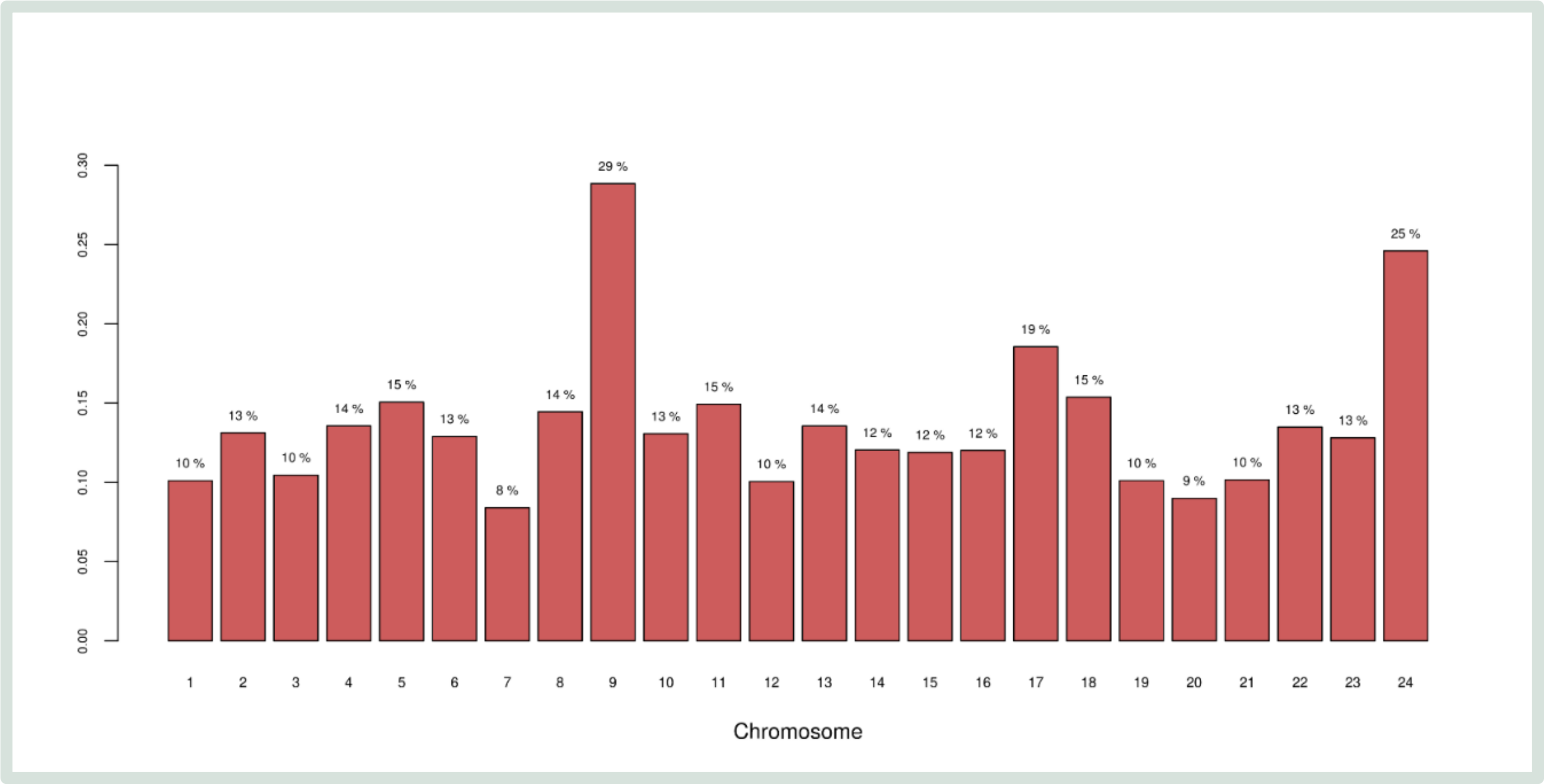


Figure 3.3.1. Percentage of disagreement per chromosome for the genotypes measured in tumor and blood. Each percentage is calculated with the number of genotypes within the chromosome with kappa < 0.8 divided by the total number of genotypes within the chromosome.

Tumors acquire frequent somatic alterations as observed with the non-concordance between tumor and germline polymorphism, which can directly impact in tumor gene expression. Thus, to integrate genomic data from both DNA sources, we excluded those tumor genotypes in agreement (wKI >0.8) between the 2 sources. As chromosome 9 was the one with the highest rate of disagreement, we restricted the study to this chromosome. The fact that we selected the chromosome 9 which is the one with the highest disagreement should be not affect the study, since we deleted those SNPs in tumor that has a wKI > 0.8 since that information was already account with the blood SNPs. The analysis was hence performed in a total number of 33,735 germline SNPs with minor allele frequency (MAF) >0.05, 8,845 tumor SNPs with MAF >0.05, 6,617 tumor CpGs located in the chromosome 9 filtered by the cross reactive probes and the polymorphic CpGs, and 717 gene expression probes with at least 20% of the individuals with expression levels >0. Three different models were applied to find the association between germline SNPs located in 1MB window (cis-relationship) with gene expression levels [Box 3].

Box 3. Models applied to the subset of chromosome 9 for the integration of 3 tumor –omics and 1 blood -omic datasets

Model 1: Global – LASSO (4-omics)

$$Gene\ Expression\ levels\ tumor_i = \alpha_1 SNPtumor_1 + \alpha_2 SNPtumor_2 + \cdots + \alpha_p SNPtumor_p + \gamma_1 CpG_1 + \gamma_2 CpG_2 + \cdots + \gamma_p CpG_p + \alpha_1 SNPblood_1 + \alpha_2 SNPblood_2 + \cdots + \alpha_p SNPblood_p ; i = 1 \dots m$$

Model 2 (2SR-MLR) & Model 3 (2SR-LASSO):

$$\begin{aligned} Gene\ Expression\ levels\ tumor_i &\\ = \alpha_1 SNPtumor_1 + \alpha_2 SNPtumor_2 + \cdots + \alpha_p SNPtumor_p &\\ + \gamma_1 CpGtumor_1 + \gamma_2 CpGtumor_2 + \cdots + \gamma_p CpGtumor_p ; i = 1 \dots m \end{aligned}$$

$$Residuals\ tumor_i = Gene\ Expression\ tumor - Gene\ \widehat{Expression}\ tumor$$

$$\begin{aligned} Residuals\ tumor_i &\\ = \alpha_1 SNPblood_1 + \alpha_2 SNPblood_2 + \cdots + \alpha_p SNPblood_p ; i &\\ = 1 \dots m \end{aligned}$$

Residuals in model 2 are obtained from MLR and for model 3 from LASSO

Model 1 (Global-LASSO) is an extension from our previous approach where germline genotypes are introduced in a linear LASSO model with the 3 tumor –omics datasets. Model 2 follows the 2SR approach applying MLR (2SR-MLR) and model 3 also follows the 2SR method but applying LASSO (2SR-LASSO) to overcome with the correlated structure between markers that cannot be made when using 2SR-MLR. Large genomic studies need to deal also with a higher number of parameters than individuals (p > n) what represents a limitation for 2SR-MLR since MLR requires n < p. To fix this issue, we used the markers selected by LASSO in 2SR-LASSO to run 2SR-MLR. The same strategy with the permutation – based MaxT method developed in (Pineda, et al. 2015) to obtain p-values and assess for Multiple Testing (MT) was applied in both Global-LASSO and 2SR-LASSO, and extended to 2SR-MLR to make the results comparable. Finally, we performed a simulation study with synthetic-data where the association between markers (SNPs tumor and blood and CpGs tumor) and gene expression levels was broken. We generated a gene expression sample with the same number of genes and individuals than in the observed data (*p*= 717, *N*= 181) considering all the genes following a normal distribution with mean (μ= 6.9) and standard deviation (σ= 0.61) extracted for the total mean and total variance from the TCGA sample.

Table 3.3.1 shows the results of the three models. After MT correction, 2SR-MLR did not selected any significant eQTL when permutation-based MaxT method is applied. In contrast, Global-LASSO selected 4,896 eQTLs involving 491 genes and 2SR-LASSO selected 266 eQTLs involving 13 genes. 2SR-MLR selected one eQTL (rs16917078-TSTD2) after applying Benjamini & Yekutieli (BY) (Benjamini & Yekutieli 2001) FDR correction to the single p-values obtained in stage 2 from the MLR model. This eQTL was also detected by Global-LASSO but not by 2SR-LASSO. As expected, Global-LASSO selected the eQTLs identified in 2SR-LASSO. The overlapping between genes and germline SNPs is represented in Figure 3.3.2 For the simulation study, 2SR-MLR did not detect any significant results with any of the MT correction methods; models Global-LASSO and 2SR-LASSO detected significant genes (6% and 3%, respectively) and germline SNPs (2% and 1%, respectively) pointing to lower proportion of false positive results yielded by 2SR-LASSO.

Table 3.3.1.Genes and germline SNPs selected by the three models in the original data (TCGA data) and the simulated dataset.

	Model 1 (Total)	Model 2 (2SR- MLR)	Model 3 (2SR-LASSO)
Original data (TCGA)			
Nº genes	491	0 (1-BY)[1]	13
Nº SNPs Germline	4,896	0 (1-BY)[1]	266
Synthetic data $N(\mu = 6,\sigma = 0.6)$			
Nº genes	43 (6%)	0	20 (3%)
Nº SNPs Germline	800 (2%)	0	566 (1%)

[1] After applying Benjamini and Yekutieli for multiple testing correction

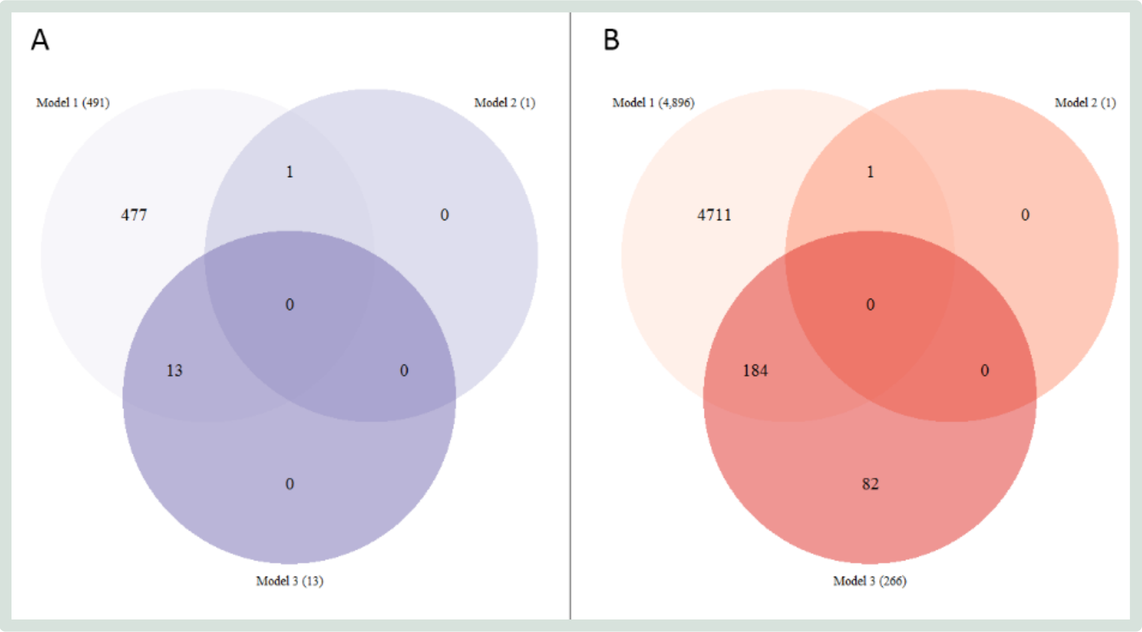


Figure 3.3.2. Overlapping of the number of genes (A) and germline SNPs (B) after MT correction using permutation – based MaxT method for model 1 and 3 and BY for model 2.

Most of the eQTLs studies addressing tumor expression have been performed without consideration of genetic and epigenetic changes in tumors (Kristensen et al. 2006; Loo et al. 2012; Chen et al. 2014). However, advances in high-throughput technologies have enabled the exploration of somatic alterations in cancer genomes(Watson et al. 2013). To perform such studies, integrative analysis are needed and consequently appropriate statistical approaches to implement them. In this study, we have shown that the multi-level integration of different source material to perform integrative eQTL assessment can be easily assess with the implementation of our approach previously proposed.

Our results confirm that the 2SR method considering MLR residuals (model 2) underestimates the associations and fail to detect eQTLs as showed in (Demissie & Cupples 2011; Che et al. 2012). This method did not produce any significant results when the permutation based MaxT method was applied to correct for MT. Moreover, when FDR was used to correct by MT only one eQTL remains significant demonstrating that the absence of significant results is not attributable to the MT method applied. In contrast, the extension of this strategy using the residuals from LASSO (model 3) produced interesting results showing an increase in power detection when using the penalized residuals, likely explained because the correlation problem between variables may be controlled by using LASSO (Tibshirani 1996). More important are the results that our strategy (model 1) provides including all the genes detected by 2SR–LASSO and the one detected with 2SR–MLR when FDR for MT correction is applied. One of the most important aspects in statistical genetics is the control for false positives, this requiring of not being too restrictive to lose valuable information (false negatives). The optimal method would provide a satisfactory balance between false positives and false negatives (Goeman & Solari 2014). In this study, to control the amount of false positives we performed a simulation study estimating that the proportion of eQTLs detected when the signal between gene expression and markers is broken was 2% (Global-LASSO) and 1% (2SR-LASSO). These percentages suggest a small proportion of signals detected that may be assumed as a rough estimation of false positives. Nonetheless, our approach detects a higher number of eQTLs than the other two approaches, thus, to be sure that the significant results were not due to an artifact of the strategy applied we run 100 times the Global-LASSO with the observed data and we applied the MaxT algorithm to assign p-values and correct for MT. With this validation, we should expect no significant results and indeed no significant results were obtained. Consequently, our approach allows to decrease the false negative rate and serve as a prioritization of interesting genes producing list of candidate genes and candidate loci to be explored in detail in future analysis. Importantly, it is that even knowing that some results may be false positives, the ones detected by the 2SR models are found also with ours and from the list of genes generated with our approach many of them have been already associated with UBC. Further details on these findings are out of the purpose of this report. The entire list of genes and associated eQTLs is provided as supplementary material (Excel S1).

Furthermore, our approach used in the Global-LASSO model is easily applicable and can accommodate any type of –omics data regardless of heterogeneity, collinearity, or number of factors in the study. Another important advantages is the reduction of computational time since the 2SR approach needs to adjust two models per gene analyzed while in our approach just one model is needed. This is a very important property when the study is extended at the genome wide level.

To conclude, eQTL IODA using different sources material may improve our knowledge in cancer risk but proper statistical methods are needed to consider the large amount of –omics data generated. We demonstrate that our approach provides a list of eQTLs that can serve for future analysis or as a prioritization to perform functional analysis. The approach is easy to apply and adaptable to any type of data being an important contribution for integrative analysis.

PART 4.

GENERAL DISCUSSON

Single –omics analysis revealed significant findings that contributed to better characterize UBC. These studies involved genetic susceptibility factors detected trough GWAS (Nathaniel Rothman et al. 2010) and EWAS (Marsit et al. 2011), or somatic DNA alterations (The Cancer Genome Atlas Research Network 2014) or gene expression signatures to predict disease progression (Kim et al. 2014), but the integration of two or more –omics considering a multi –omics approach will decipher further interrogations not covered by a single –omics data type analysis. However, dealing with such amount of data, coming frequently from high-throughput technologies, impose many challenges than need to be addressed appropriately (Hamid et al. 2009; Chadeau-Hyam et al. 2013; Kristensen et al. 2014; Ritchie et al. 2015). Having this scenario in mind, our objective was to dissect and fix some of the methodological challenges posed by –omics data integration. The work described in part 3 of this thesis shows that applying appropriate –omics integrative statistical strategies, sound biological insights in the complexity of UBC are discovered.

The first step in any integrative –omics approach implementation is to conduct data quality control and assess data scale and dimensionality for each dataset component. A detailed single analysis of each –omics component was described in part 2. Regarding genomics, we highlight the differences found when measure them in blood and tumor (Figure 2.1.2) due to the somatic mutations acquire in the tumor. While it is known that tumor DNA is affected by somatic mutations, to our knowledge, no articles are published assessing it at SNP level. Only few studies showed that variants on pharmacogenetic genes are not affected by differences between germline and tumoral SNPs (McWhinney & McLeod 2009; Weiss et al. 2007; Marsh et al. 2005). We found that in bladder cancer, the blood-tumor genotypes differences are frequent in terms of disagreement percentages, especially for chromosome 9 (25%), Y (13%), 17 (7%), 8 (5%) and 11 (5%). The differences in chromosome 9 are explained due to the deletion of both arms of this chromosome that occurs early during the urothelial tumorigenesis (Wu 2005). Also deletions in the short arms of chromosome 8 and 11 are associated with tumor progression (Wu 2005). We also performed this study using the TCGA data (Figure 3.3.1) where the blood-tumor genotypes differences in chromosome 9 (29%), Y (25%) and 17 (19%) are also observed in MIBC, differences in chromosome Y and 17 being even larger. Importantly, these results suggest that there is enough disagreement between both measures to consider this data as two different –omics data sets. In the epigenomics single analysis, we highlight the differences found between autosomal chromosomes and X-chromosome (Figure 2.2.2) supporting previous evidences (Bell et al. 2011) and is explained by one of the X-chromosome inactivation in females through methylation mechanisms. It is also important to consider the differences observed between the distributions of the *β*- and the *M*-values. While the former follow a beta distribution in the autosomal chromosomes taking values from 0 to 1, the *later* follow a bimodal distribution taking values from -∞ to +∞. This last measure achieves the homocedastic characteristic which is a needed assumption in most of the statistical methods including the ones used in this thesis. Gene expression data follow a normal distribution as showed in Figure 2.3.4 facilitating the use of parametric statistical methods for its analysis.

This thesis represents an important advancement in –omics data integration not only because of the number of –omics data sets that have been combined (3 –omics data in tumor plus 1 –omics data measured in a different source: blood) but mainly because of the innovative analytical methods we propose and apply in their integration. First, a framework build upon a multi-staged strategy analyzing all the possible pairwise relationships between genomics, epigenomics and trasncriptomics was proposed and applied using tumor data from individuals with UBC included in the EPICURO pilot phase. The framework (Figure 3.1.1) proposes 4 consecutive steps starting with the preprocessing and QC of each –omics dataset aforementioned (Step 1). In Step 2, the pairwise analysis between genomics and transcriptomics is assessed showing interesting correlations between DNA methylation and gene expression. In Step 3, the eQTL and metQTL analysis is performed with the correlated expression-methylation pairs ($\rho \geq |0.7|$) obtained in Step 2. Finally, in the Step 4, the combination of the pairwise analysis in an integrative way is assessed showing interesting results: a 10% of enrichment of commonly associated genetic variants (49,708 SNPs) with both, the eQTLs and methQTLs, a total of 1,469 *trans* triple relationships (SNP – CpG – Expression), being 19 of the pairs involved in a *cis* relationship and regions with hotspots (Figure 3.1.3) showing some important biological relationships.

This framework, in comparison with other integrative analysis using the same three –omics data types, integrates step by step the pairwise relationships resulting in a final combination showing triples relationships that are lastly explored in more detail. For example, in comparison with the analysis showed by Gibbs et al. (2010) and Bell et al. (2011), they both perform first, the eQTL and methQTL analysis and then check for the overlapping SNPs between both analyses. But, they do not show a combination between the three types of data as we do, although they do not restrict the eQTLs and methQTLs to the correlated expression-methylation pairs as we do. The reason why we restricted the analysis is because we were interesting in provide a framework to integrate the 3-omics over the specific pairwise relationship analysis. Another example is the study of Wagner et al. (2014) that perform an integrative analysis, first analyzing eQTLs and methQTLs and then looking at the correlation between gene expression and DNA methylation to find the overlapping between the three data types. They used genes rather than CpGs or SNPs as the primary unit of interest for the overlapping, while we used the specific SNPs and CpGs for the overlapping. Also they only considered *cis*-relationships while we considered both, *cis* and *trans*. It is well established that gene expression levels are controlled by a combination of *cis* and *trans*-acting regulators (Cheung & Spielman 2009). Therefore, the study of *trans* relationships it is very important to consider in the biological complexity of cancer. In the case of Zhang et al. (2010), Drong et al. (2013) and Olsson et al. (2014), they focus the analysis in the methQTL analysis and then, with the SNPs that were significant, perform subsequent eQTLs analysis. This limit the study to differentially methylated regions of interest but not at genome wide level. One study that could be considered that integrates all the pairwise analysis at genome wide level as we do, is Van Eijk et al. (2012) that first, obtain the correlation between DNA methylation and gene expression levels and then perform eQTL and methQTLs analysis. They perform an accurate multi-staged analysis where they assess with the final subset of significant relationships a causal relationships study between the three types of data while we assess specific triples relationships to find patterns of relationships and interesting regions (hotspots). In conclusion, our framework not only accomplish pairwise analysis, but also provides the integration of the three –omics data combining the significant results from the pairwise analysis as showed in Step 4.

While the previous approach allows advancing in the integration process by dissecting the pairwise relationships between the three –omics data step by step, it does not permit the assessment of all potential associations and adjustments among the whole set of markers. To advance in this more comprehensive regard, all –omics data were combined in a large input matrix using penalized regression methods and we adapted a permutation –based MaxT method to correct for multiple testing and provide p-values. In the introduction, we commented about the challenges –omics integrative analysis imposes (heterogeneity, dimensionality, n << p, correlation, interpretation, replication and validation) which the majority are addressed in this work by using two types of penalized regression methods, LASSO and ENET. We also commented on the pitfalls of these two methods regarding the absence to assess significance and consequently correct by MT. With this work we are able to deal with these methodological challenges. LASSO and ENET are flexible strategies and can be applied even if the variables have different scale. They are variable selection methods and consequently they can deal with high dimensionality. They can be applied when the sample size is smaller than the number of parameters and also when the variables are correlated. Besides, with the permutation –based MaxT method, we are able to obtain p-values and correct by MT capturing only the significant models that have passed the threshold for MT, thus they produce sparse results to be interpretable. After applying the workflow proposed in material&methods (Figure 3.2.1), the expression of 48 genes was found significantly associated with several SNPs and CpGs in *cis*-acting effect. Interesting are the differences found by applying the two methods LASSO and ENET to the three models, SNP model, CPG model and Global model (SNP+CPG). The reason is due to the selection process of both methods. ENET tends to select nets of variables that are highly correlated while LASSO select just one representative variable of the net. Also differences were found when applying the three models. The main difference between models regards to the genes identified only when combining together SNPs and CpGs likely explained because of the addition of more information when integrating, a fact that further supports the importance of integrating –omics data to discover hidden information.

As we highlighted in the introduction, the validation of these complexities is a big challenge and new ideas for replicating are needed. Typically, replication studies have been based in reproducing the same relationship between the marker and the trait in the study, but this is very difficult in the –omics field due to the lack of a similar population with same –omics data sets measured in the same platforms. The TCGA has generated data for individuals with UBC though they are MIBC and all the –omics platforms used are different from ours. Because of this massive heterogeneity, a replication study at gene level was considered in this work by wandering whether the genes selected in our study considering 1Mb window up- and downstream were also selected in the TCGA series using the same analytical strategy. Importantly, we found that the 75% of the identified genes (64% of the models) were replicated supporting our approach. In addition, to validate our approach, we checked the literature and we found that several genes were previously shown to be important in bladder cancer (*KRT20, IGF2, CTSE, ANXA10* and *CRH*) (Mengual et al. 2010). Furthermore, an enrichment analysis was also performed for biological validation and two clusters were found, the “EF hand and calcium ion binding” that includes two important genes (*S100A9* and *S100A2)* being previously associated with UBC (Yao et al. 2007; Dokun et al. 2008; Minami et al. 2010). This approach demonstrates that the integration of multiple –omics data allows the identification of hidden information missed when only one –omics is studied. Previous advantages have been performed to combine data matrices for each sample into one large input matrix to perform integrative analysis (Fridley et al. 2012; Mankoo et al. 2011; Kim et al. 2013). They all applied different and valid approaches to integrate omics data, but they all predict one phenotype, such as survival risk or drug cytotoxicity. Thus, they all have the limitation to integrate –omics data considering more than one output variable to be estimated as we do considering all the probes in the gene expression array as independent output variables. Our approach is flexible and easy to implement in different –omics data and disease models as well as adaptable to consider interaction terms in the model.

The developed approach abovementioned, we also used to perform an integrative eQTL analysis considering two types of source material (blood and tumor). This approach was compared with the 2SR strategy using MLR for eQTL integrative analysis conducted in other studies (Li et al. 2014; Li et al. 2013). Our results showed no significant eQTLs when using 2SR approach according to the evaluation performed of this method by Demissie & Cupples (2011) and Che et al. (2012) confirming the underestimation and loss of power of the results. We proposed extending this method applying LASSO and using its residuals to avoid the problem that MLR have with correlated variables. By doing this, we found significant results suggesting an increase of power of the method. Considerable are the results from our approach that increase remarkably the number of significant results including the ones from the other approaches. We are aware that false positives may be an issue but one of the most important aspects in statistical genetics with difficult solution until now, is how to control by MT while not being too restrictive (Goeman & Solari 2014). Assuming that a proportion of false positives may occur, our approach can serve as a prioritization of interesting results producing a list of candidate genes and loci to be explored in detail in future analyses with the advantages that it can be applied easily accommodating different –omics, it deals with heterogeneity and collinearity or the number of factors, and it reduces computational time.

Overall, this work has strengths and limitations that should be mentioned. The sample size is a limitation in –omics integrative studies and certainly in this thesis. Nevertheless, one of the strengths for the multi-staged approach responds to the fact that the overlap for the samples available between pairs (epigenomics – transcriptomics = 31, genomics-epigenomics = 46, genomics – transcriptomics = 27) is higher than the overlap between triplets (genomics-epigenomics-transcriptomics = 27). In the multi-staged approach, pairwise combinations are applied to obtain the integrative results using the specific sample in each step increasing them the size. Despite this issue, the problem of the n << p obligates to perform a one marker analysis per model which always is a limitation when correcting by MT. In contrast, LASSO and ENET deal with this issue although sample size is very small to yield strong biological conclusions. This work needs to be seen as a methodological contribution that adds different –omics data integrative approaches to be applied in bigger sample sizes. In this regards, interesting is that when applying LASSO & permutation –based MaxT method to perform the integrative eQTL analysis with a bigger sample size (181 individuals) as showed in chapter 3 of PART3, the number of eQTLs detected increase remarkably. One potential explanation could be the increase in power due to the increase of sample size, although we need to be cautious when comparing these two studies because the first one applied to EPICURO data, the gene expression levels are measured with microarray technology, while the second one applied to TCGA data, the expression levels are measured with RNA sequencing. Microarray data follows a normal distribution and does not contain zeros giving a relative value of gene expression levels while RNA sequencing gives the absolute values of gene expression levels with zero value when the gene is not expressed that, in fact, makes the genes not being normally distributed.

For this thesis, no filter by LD in SNPs or grouping CpGs in a CpG island was considered. The reason is that by performing data reduction the causal markers may be filtered and thereby the true association may be lost. Also, some of the statistical methods that perform data reduction, such as PC or FA, convert the markers into linear combinations of the original ones making difficult the understanding of the original values. In the first approach from our work, as no filter was performed, to avoid the problem of the possible correlation between p-values due to the correlation between variables, Benjamini and Yekuteili FDR (Benjamini & Yekutieli 2001) MT approach was applied since this method can be used under general dependence between tests through the null distribution resampling. In the second approach, LASSO and ENET can deal directly with collinearity and high dimensionality and hence no filtering was needed.

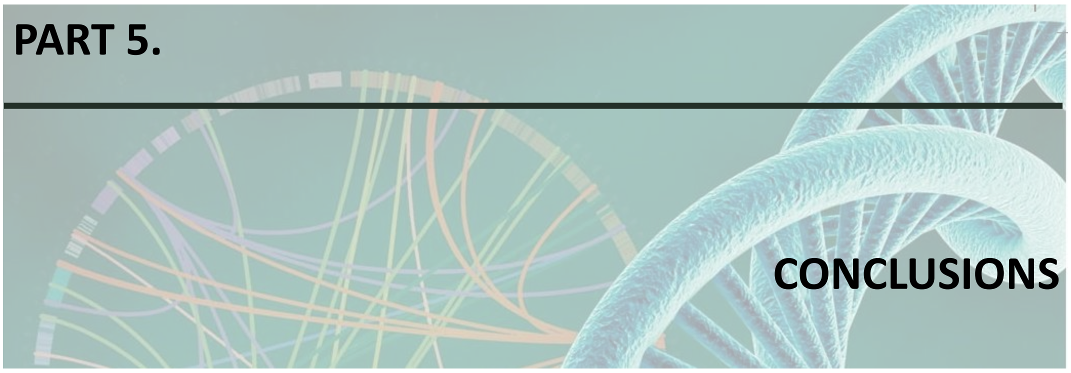
It is very important to take into consideration in which type of data the integrative analysis is performed. Tissue selection is an important factor to consider for eQTL analysis and consequently for any genomics integrative analysis. For instance, the effect of a SNP on a transcript may only be revealed in a tissue-specific manner. Studies that have been performed in individuals with breast cancer showed an association between a specific locus and FGFR2 in tumor tissue (Meyer et al. 2008), but the association was not showed in normal tissue (Seo et al. 2013). For this reason, in this thesis we have been cautious with the type of data considered in each analysis. For the two first approaches we used tumor tissue which gave us the opportunity to study in detail the regulation in the tumor setting, and in the last work, tumor tissue was considered to estimate the somatic alterations before assessing the risk of loci in blood samples associated with tumoral gene expression. On the other hand, a potential limitation derived from the use of tumor samples, since they can be contaminated by stroma cells, such as for example one of the genes found significant in the second approach (*IGJ*) was not expressed in UBC cell lines (Earl et al. 2015).

An important consideration in all –omics studies is how to validate results both at the biological interpretation level of the list of results generated and on at the replication level to determine whether the results are more likely to be true positives than false positives. In this thesis, a biological validation-like was performed looking for the existence of published results that would help in the interpretation. In addition, an enrichment analysis was performed with DAVID bioinformatics tool to find clusters within our final results showing interesting biological information that further support our approaches. Nevertheless the biological interpretation is still a challenge to deal with because of the lack of biological information at present. In regards with the replication, the ideal scenario would be to replicate the results in an independent data set, but the lack of a comparable and independent UBC patient series with the three –omics data available makes this difficult. As pointing in the introduction, new ideas of validating, replicating and interpreting this type of results are needed.

The future directions for this work are based on the generation of high-throughput –omics data in very large sample sizes that will be soon possible due to the decrease in cost and the amount of public data/samples available. Playing with these voluminous dataset, we will be able to improve our statistical strategies to integrate them. A consequent issue of this is the computational storage and the computational time from which to find the best computational strategies will be needed to work with big data. This is a field were the statistics, the biology and the informatics are crossed and it is clear that we will require of a strong knowledge from different skills and therefore multidisciplinary teams and collaborative work will be needed. But in this equation, the role of the epidemiology is lacking. Epidemiologist aim to integrate this massive amount of –omics data generated with the non–omics information coming from other sources, such as questioners, candidate markers, etc. Adding this information to the integrative type of analysis that is shown in this thesis will lead to the building of better predictive risk models. But the integration of these non-omics type of data poses other methodological challenges that all together need to face. As an attempt to approach the –omics and non-omics data integration, (López de Maturana et al. 2015) has submitted a review on this regard where I also collaborate.

Another future endeavor is to model the causal relationships between the omics data to link them to the phenotype outcome. Among the proposals, is the one used in (Olsson et al. 2014) where they used a causal inference test (CIT) (Millstein et al. 2009) to model the causal relationship between genotype, DNA methylation and phenotypic outcome, or the one used in (Van Eijk et al. 2012) were they infer the directionality in the relationship between genetic variants, methylation and expression with the local edge oriented (LEO) scores based on structural equation models (Aten et al. 2008).

To finish, integrative –omics analysis are needed to find missing, hidden or unreliable information. For instance, examining genetic and epigenetic changes in gene expression pattern improves the identification of causal changes that lead to disease phenotypes. Evolving statistical procedures that operates in the integration of more than one single -omics will be critical to extract information related to complex diseases, as research goes beyond a single – omics focus. The early success of the approaches described in this thesis to better characterize bladder cancer, will significantly enhance the identification of key drivers of disease beyond what could be achieved by one –omics assessment alone.



PART 5.

CONCLUSIONS

1. We demonstrated that appropriate integrative –omics data analysis allows to identify hidden genomics mechanisms not observed when analyzing a single –omics data type by implementing
    a. A multi-stage framework to analyses pairwise combinations;
    b. A multidimensional approach using penalized regression methods with a permutation-based MaxT approach to integrate >2 –omics data in the same model; and
    c. An extension of the previous approach to integrate >2 –omics data from different levels (tumor and blood).

2. The propose multi-stage framework not only accomplished pairwise analysis, but also allowed the integration of the three –omics data by combining the significant results from each pairwise analysis. Applying the integrative framework to the SBC/EPICURO Study data, relevant integrative results were found:
    a. 10% of enrichment of common genetic variants (49,708 SNPs) associated with both gene expression (eQTLs) and methylation (methQTLs);
    b. 1,469 *trans-acting* triple significant relationships (SNP – CpG – expression);
    c. 19 pairs involved in a *cis-acting* relationship, and
    d. "hotspots" genetic regions suggesting predominant relationships.

3. The multidimensional approach using penalized regression methods (LASSO and ENET) allowed us to deal with some of the main challenges –omics data integration impose (heterogeneity, dimensionality, n << p and multicolinearity). In addition, in combination with a permutation –based MaxT method, we were able to assess the statistical significance of the models and to provide a multiple testing corrected p-value for each association.

4. Applying the penalized regression & permutation –based MaxT strategy to the SBC/EPICURO Study data, we found a list of 48 genes differently expressed according to several SNP genotypes and CpGs levels in *cis-acting* relationship. 75% of the identified genes were replicated in an independent data set (TCGA Consortium data) despite of the important heterogeneity between data sets. Furthermore, we provided biological interpretation of the results with an enrichment analysis highlighting the "EF hand and calcium ion binding" cluster involving two genes (*S100A9* and *S100A2)* previously associated with UBC.

5. Using the penalized regression & permutation –based MaxT strategy to integrate 4-omics data from different levels (3-omics in tumor, 1-omics in blood), we were able to perform an integrative eQTL analysis with TCGA Consortium data. This strategy permited us to adjust for potential alterations in tumor when assessing the association of germline SNPs with tumor gene expression. We demonstrated that our strategy found the same results than the 2SR-LASSO approach (an extension of the 2SR-MLR), in addition to increase a higher number of significant eQTLs, suggesting an increase in statistical power.

1. Hemos demostrado que la aplicación de análisis de integración de varios datos –ómicos permite identificar mecanismos genómicos ocultos no observados con el análisis de un solo tipo de dato –ómico. Ello ha sido posible implementando:
    a. Un marco de análisis en varias fases para combinar datos –ómicos por parejas;
    b. Un enfoque multidimensional utilizando métodos de regresión penalizada con un método basado en permutaciones MaxT para integrar >2 datos –ómicos en el mismo modelo; y
    c. Una extensión del enfoque anterior para integrar >2 datos –ómicos de diferentes niveles (tumor y sangre).

2. El marco de análisis en varias fases no sólo analiza las combinaciones por pares, sino que también proporciona la integración de los tres datos –ómicos combinando los resultados significativos del análisis por pares. Aplicando este marco a los datos del estudio SBC/EPICURO, se encontraron los siguientes resultados relevantes:
    a. Un enriquecimiento del 10% en 49,708 variantes genéticas comunes asociadas con ambos expresión del gen (eQTLs) y metilación de ADN (methQTLs);
    b. Un total de 1,469 relaciones triples en *trans-acting* (SNP - CpG - Expresión); y
    c. Regiones genéticas llamadas "puntos calientes" que concentran relaciones con otras regiones del genoma.

3. El enfoque multidimensional usando métodos de regresión penalizada (LASSO y ENET) nos permitió hacer frente a algunos de los principales retos que los datos –ómicos imponen (heterogeneidad, dimensionalidad, n << p y multicolinealidad). Además, en combinación con un método basado en permutaciones MaxT, fuimos capaces de evaluar la bondad de ajuste de los modelos y proporcionar un p-valor corregido por comparaciones múltiples.

4. Aplicando métodos de regresión penalizada y un método basado en permutaciones MaxT a los datos del estudio SBC/EPICURO, encontramos una lista de 48 genes asociados a varios polimorfismos y niveles de metilación del ADN en relaciones *cis-acting*. Además, replicamos el 75% de los genes identificados con una base de datos independiente (datos del consorcio TCGA) a pesar de la gran heterogeneidad entre estas dos bases de datos. Además, proporcionamos una interpretación biológica de los resultados mediante un análisis de enriquecimiento destacando el "EF-hand and calcium ion binding" como un grupo que contiene a dos genes (*S100A9* y *S100A2*) previamente asociados con cáncer de vejiga.

5. Usando métodos de regresión penalizada y un método basado en permutaciones MaxT para integrar 4 tipos de datos -ómicos procedentes de material diferente (3-ómicas en tumor, 1-ómica en sangre) a los datos del Consorcio TCGA, hemos sido capaces de realizar un análisis de integración de eQTLs ajustando primero las posibles alteraciones -ómicas en el tumor y analizando después los SNPs en línea germinal asociados a los niveles de expresión del gen en estudio. Hemos demostrado que usando nuestra estrategia, no sólo encontramos los mismos resultados que con la regresión en 2 etapas aplicando LASSO (una extensión de la regresión en 2 etapas aplicando regresión múltiple), sino que también aumentamos el número de eQTLs significativas sugiriendo un incremento de poder estadístico.

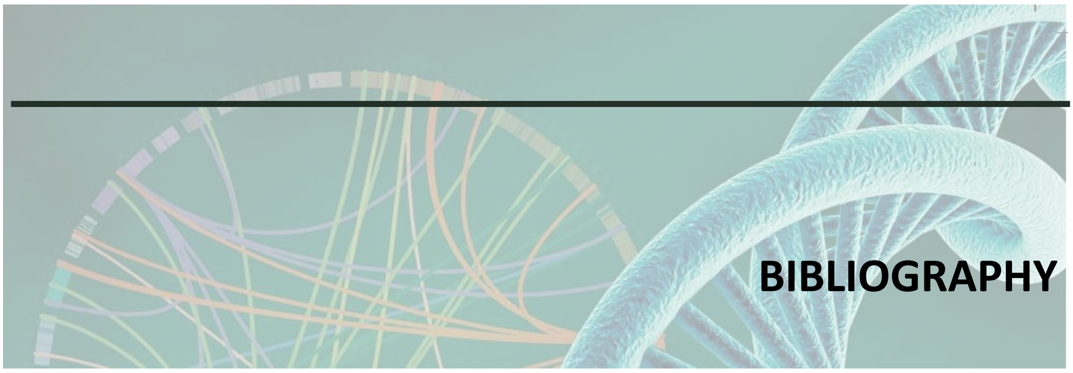
1. Nous avons montré que l'integration de données –omiques nous permettait d'identifier des mécanismes génomiques cachés non observés dans l'analyse d'un seul type de donnée -ómique via:
    a. Une stratégie multi-étapes qui propose un cadre en utilisant les combinaisons de paires ;
    b. Une stratégie multidimensionnelle, utilisant des méthodes de régressions pénalisées avec des permutations MaxT pour intégrer >2 données -omiques dans le même modèle ; et
    c. Une extension de celui-ci pour intégrer >2 données -omiques issues de matériaux différent (tumeur et sang).

2. Le cadre proposé permet non seulement l'analyse des combinaisons de paires, mais fournit également l'intégration des trois données -omiques en combinant les résultats d'analyse des paires significatives. L'application de ce cadre aux données de l'étude SBC/EPICURO, a conduit à des résultats significatifs:
    a. Un enrichissement de 10% des 49,708 variantes génétiques communément associés conjointement aux eQTLs et methQTLs ;
    b. Un total de 1,469 relations *trans* triples (SNP - CPG - expression) ;
    c. Certaines régions génétiques "points chauds" qui suggèrent des concentrations des relations avec d'autres régions du génome.

3. Grace à une approche multidimensionnelle utilisant des méthodes de régression pénalisés (LASSO et ENET), nous avons pu relever certains défis majeurs imposés par les données -omiques (hétérogénéité, dimensionnalité, n << p multicolinéarité). Par ailleurs, en combinaison avec une méthode basée sur des permutations MaxT, nous avons été en mesure d'évaluer la qualité de l'ajustement des modèles et de fournir une valeur de p corrigée pour les comparaisons multiples pour chaque association.

4. En appliquant des méthodes de régressions pénalisées et une méthode basée sur des permutations MaxT aux données de l'étude SBC/EPICURO, nous avons trouvé une liste de 48 gènes exprimés différentiellement en fonction des polymorphismes et des différents niveaux de méthylation de l'ADN dans une relation *cis-acting*. Nous reproduisons 75% des gènes dans une base de données distincte (données du Consortium TCGA) malgré la grande hétérogénéité entre ces deux bases de données. Nous fournissons également une interprétation biologique des résultats dans une analyse d'enrichissement pointant le groupe " EF-hand and calcium ion binding" impliquant deux gènes (*S100A9* et *S100A2*) précédemment associés avec le cancer de la vessie.

5. L'application des méthodes de régression pénalisée et d'une méthode basée sur des permutations MaxT pour intégrer quatre -omiques issus de différents matériaux biologiques (3-omiques tumeur, 1-omique sang) sur les données d'étude TCGA, a permis d'effectuer une analyse d'intégration des eQTLs. Les modèles ont été ajustés aux possibles altérations dans la tumeur permettant d'identifier des variants germinaux associés aux niveaux d'expression des gènes à l'étude. Nous avons démontré que l'utilisation de notre stratégie permettait non seulement d'identifier les mêmes résultats qu'avec la régression en deux étapes utilisant le LASSO (une extension de la régression en deux étapes à l'aide de la régression multiple), mais aussi d'augmenter le nombre des eQTLs significatifs suggérant une puissance statistique accrue.



BIBLIOGRAPHY

Aten, J.E. et al., 2008. Using genetic markers to orient the edges in quantitative trait networks: the NEO software. *BMC Systems Biology*, 2, p.34.

Ayers, K.L. & Cordell, H.J., 2010. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiology*, 34(8), pp.879–91.

Banovich, N.E. et al., 2014. Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genetics*, 10(9), p.e1004663.

Bell, C.G. et al., 2010. Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the FTO type 2 diabetes and obesity susceptibility locus. *PLoS One*, 5(11), p.e14040.

Bell, J.T. et al., 2011. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology*, 12(1), p.R10.

Benjamini, Y. & Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series b (Methodological)*, 57(1), p.11.

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

Bibikova, M. et al., 2009. Genome-wide DNA methylation profiling using Infinium(R) assay. *Epigenomics*, 1(1), pp.177–200.

Bickmore, W.A. & van Steensel, B., 2013. Genome architecture: domain organization of interphase chromosomes. *Cell*, 152(6), pp.1270–84.

Bird, A., 2002. DNA methylation patterns and epigenetic memory. *Genes & Development*, 16(1), pp.6–21.

Browning, S.R. & Browning, B.L., 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, 81(5), pp.1084–97.

Bryois, J. et al., 2014. Cis and trans effects of human genomic variants on gene expression. C. D. Brown, ed. *PLoS Genetics*, 10(7), p.e1004461.

Bush, W.S., Dudek, S.M. & Ritchie, M.D., 2009. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp.368–79.

Byun, H.-M. et al., 2007. Examination of IGF2 and H19 loss of imprinting in bladder cancer. *Cancer Research*, 67(22), pp.10753–8.

Carrel, L. & Willard, H.F., 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature*, 434(7031), pp.400–4.

Chadeau-Hyam, M. et al., 2013. Deciphering the complex: methodological overview of statistical models to derive OMICS-based biomarkers. *Environ Mol Mutagen*, 54(7), pp.542–557.

Che, R., Motsinger-Reif, A.A. & Brown, C.C., 2012. Loss of power in two-stage residual-outcome regression analysis in genetic association studies. *Genetic Epidemiology*, 36(8), pp.890–4.

Chen, L.S. et al., 2010. Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *American Journal of Human Genetics*, 86(6), pp.860–71.

Chen, Q.-R. et al., 2014. Systematic genetic analysis identifies Cis-eQTL target genes associated with glioblastoma patient survival. *PloS One*, 9(8), p.e105393.

Chen, Y. et al., 2011. Sequence overlap between autosomal and sex-linked probes on the Illumina HumanMethylation27 microarray. *Genomics*, 97(4), pp.214–22.

Cheung, V.G. & Spielman, R.S., 2009. Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nature Reviews. Genetics*, 10(9), pp.595–604.

Cho, S. et al., 2010. Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis. *Annals of Human Genetics*, 74(5), pp.416–28.

Choi, W. et al., 2014. Intrinsic basal and luminal subtypes of muscle-invasive bladder cancer. *Nature Reviews. Urology*, 11(7), pp.400–10.

Demissie, S. & Cupples, L.A., 2011. Bias due to two-stage residual-outcome regression analysis in genetic association studies. *Genetic Epidemiology*, 35(7), pp.592–6.

Dennis, G. et al., 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*, 4(5), p.P3.

Dokun, O.Y. et al., 2008. Relationship of SNCG, S100A4, S100A9 and LCN2 gene expression and DNA methylation in bladder cancer. *International journal of cancer. Journal International of Cancer*, 123(12), pp.2798–807.

Drong, A.W. et al., 2013. The presence of methylation quantitative trait loci indicates a direct genetic influence on the level of DNA methylation in adipose tissue. *PLoS One*, 8(2), p.e55923.

Du, P. et al., 2010. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11, p.587.

Earl, J. et al., 2015. The UBC-40 Urothelial Bladder Cancer cell line index: a genomic resource for functional studies. *BMC Genomics*, 16(1), p.403.

Van Eijk, K.R. et al., 2012. Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics*, 13, p.636.

Encode Project Consortium, 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, N.Y.)*, 306(5696), pp.636–40.

Escofier, B. & Pagès, J., 1994. Multiple factor analysis (AFMULT package). *Computational Statistics & Data Analysis*, 18(1), pp.121–140.

Esteller, M., 2008. Epigenetics in cancer. *N Engl J Med*, 358(11), pp.1148–1159.

Ferlay, J. et al., 2013. *GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11*, Lyon, France: International Agency for Research on Cancer.

Fredriksson, N.J. et al., 2014. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nature Genetics*, 46(12), pp.1258–63.

Fridley, B.L. et al., 2012. A Bayesian integrative genomic model for pathway analysis of complex traits. *Genetic Epidemiology*, 36(4), pp.352–9.

Fu, Y.-P. et al., 2012. Common genetic variants in the PSCA gene influence gene expression and bladder cancer risk. *Proceedings of the National Academy of Sciences of the United States of America*, 109(13), pp.4974–9.

García-Closas, M. et al., 2006. NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses. *Lancet*, 366(9486), pp.649–59.

Gautier, L. et al., 2004. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3), pp.307–315.

Ge, Y., Dudoit, S. & Speed, T.P., 2003. Resampling-based multiple testing for microarray data analysis. *Test*, 12(1), pp.1–77.

Gibbs, J.R. et al., 2010. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genetics*, 6(5), p.e1000952.

Gilbert, P.M. et al., 2010. HOXA9 regulates BRCA1 expression to modulate human breast tumor phenotype. *The Journal of Clinical investigation*, 120(5), pp.1535–50.

Goeman, J.J. & Solari, A., 2014. Multiple hypothesis testing in genomics. *Statistics in Medicine*, 33(11), pp.1946–78.

Greenawalt, D.M. et al., 2012. Integrating genetic association, genetics of gene expression, and single nucleotide polymorphism set analysis to identify susceptibility Loci for type 2 diabetes mellitus. *American Journal of Epidemiology*, 176(5), pp.423–430.

Guerrero-Preston, R. et al., 2011. NID2 and HOXA9 promoter hypermethylation as biomarkers for prevention and early detection in oral cavity squamous cell carcinoma tissues and saliva. *Cancer Prevention Research (Philadelphia, Pa.)*, 4(7), pp.1061–72.

Hamid, J.S. et al., 2009. Data integration in genetics and genomics: methods and challenges. *Hum Genomics Proteomics*, 2009.

Hellman, A. & Chess, A., 2007. Gene body-specific methylation on the active X chromosome. *Science (New York, N.Y.)*, 315(5815), pp.1141–3.

Heyn, H. et al., 2014. Linkage of DNA methylation quantitative trait Loci to human cancer risk. *Cell Rep*, 7(2), pp.331–338.

Hoerl, A.E. & Kennard, R.W., 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1), pp.55–67.

Huang, D.W., Sherman, B.T. & Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1), pp.44–57.

Hui Zou, T.H., 2005. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, 67, pp.301–320.

Human Genome Sequencing ConsortiumInternational, 2004. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), pp.931–45.

Hunter, D.J. et al., 2007. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics*, 39(7), pp.870–4.

Ibragimova, I. et al., 2014. A global profile of gene promoter methylation in treatment-naïve urothelial cancer. *Epigenetics : official journal of the DNA Methylation Society*, 9(5), pp.760–73.

Irizarry, R.A. et al., 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)*, 4(2), pp.249–64.

Irizarry, R.A. et al., 2009. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Genetics*, 41(2), pp.178–86.

Jerome Firedman; Trevor Hastie; Rob Tibshirani, 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1).

Jollife, I.I., 2002. *Principal Component Analysis*, Springer Series in Statistics.

Jones, P.A., 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature reviews. Genetics*, 13(7), pp.484–92.

Kanwal, R. & Gupta, S., 2012. Epigenetic modifications in cancer. *Clinical Genetics*, 81(4), pp.303–11.

Karagiannis, G.S., Pavlou, M.P. & Diamandis, E.P., 2010. Cancer secretomics reveal pathophysiological pathways in cancer molecular oncology. *Molecular Oncology*, 4(6), pp.496–510.

Kauffmann, A., Gentleman, R. & Huber, W., 2009. arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25(3), pp.415–416.

Kim, D. et al., 2013. ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. *BioData Mining*, 6(1), p.23.

Kim, W.T. et al., 2014. S100A9 and EGFR gene signatures predict disease progression in muscle invasive bladder cancer patients after chemotherapy. *Annals of oncology : Official Journal of the European Society for Medical Oncology / ESMO*, 25(5), pp.974–9.

Knowles, M.A. & Hurst, C.D., 2014. Molecular biology of bladder cancer: new insights into pathogenesis and clinical diversity. *Nature Reviews Cancer*, 15(1), pp.25–41.

Kristensen, V.N. et al., 2006. Genetic variation in putative regulatory loci controlling gene expression in breast cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 103(20), pp.7735–40.

Kristensen, V.N. et al., 2014. Principles and methods of integrative genomic analyses in cancer. *Nat Rev Cancer*, 14(5), pp.299–313.

Krzywinski, M. et al., 2009. Circos: an information aesthetic for comparative genomics. *Genome Research*, 19(9), pp.1639–45.

Leung, D. et al., 2015. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*, 518(7539), pp.350–354.

Li, Q. et al., 2014. Expression QTL-based analyses reveal candidate causal genes and loci across five tumor types. *Human Molecular Genetics*, 23(19), pp.5294–302.

Li, Q. et al., 2013. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell*, 152(3), pp.633–641.

Lichtenstein, P. et al., 2000. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *The New England Journal of Medicine*, 343(2), pp.78–85.

Liu, Y. et al., 2013. Methylomics of gene expression in human monocytes. *Hum Mol Genet*, 22(24), pp.5065–5074.

Loo, L.W.M. et al., 2012. cis-Expression QTL analysis of established colorectal cancer risk variants in colon tumors and adjacent normal tissue. *PloS One*, 7(2), p.e30477.

López de Maturana, E. et al., 2015. Towards the integration of omics data in epidemiological studies: still a “long and winding road.” *Genetic Epidemiology*.

López-Knowles, E. et al., 2006. PIK3CA mutations are an early genetic alteration associated with FGFR3 mutations in superficial papillary bladder tumors. *Cancer Research*, 66(15), pp.7401–4.

Lu, L.J. et al., 2005. Assessing the limits of genomic data integration for predicting protein networks. *Genome Research*, 15(7), pp.945–53.

Lyon, M.F., 1961. Gene action in the X-chromosome of the mouse (Mus musculus L.). *Nature*, 190, pp.372–3.

MacDonald JW, hugene10sttranscriptcluster.db: Affymetrix hugene10 annotation data (chip hugene10sttranscriptcluster). R package version 8.3.1.

Makishima, H. et al., 2013. Somatic SETBP1 mutations in myeloid malignancies. *Nature genetics*, 45(8), pp.942–6.

Malats, N. & Real, F.X., 2015. Epidemiology of bladder cancer. *Hematology/Oncology Clinics of North America*, 29(2), pp.177–89, vii.

Mankoo, P.K. et al., 2011. Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. S. Deb, ed. *PloS One*, 6(11), p.e24709.

Marsh, S. et al., 2005. Concordance of pharmacogenetic markers in germline and colorectal tumor DNA. *Pharmacogenomics*, 6(8), pp.873–7.

Marsit, C.J. et al., 2011. DNA Methylation Array Analysis Identifies Profiles of Blood-Derived DNA Methylation Associated With Bladder Cancer. *Journal of Clinical Oncology*, 29(9), pp.1133–1139.

McWhinney, S.R. & McLeod, H.L., 2009. Using germline genotype in cancer pharmacogenetic studies. *Pharmacogenomics*, 10(3), pp.489–93.

Meinshausen, N., Maathuis, M.H. & Bühlmann, P., 2011. Asymptotic optimality of the Westfall–Young permutation procedure for multiple testing under dependence. *The Annals of Statistics*, 39(6), pp.3369–3391.

Mengual, L. et al., 2010. Gene expression signature in urine for diagnosing and assessing aggressiveness of bladder urothelial carcinoma. *Clinical cancer research : An Official Journal of the American Association for Cancer Research*, 16(9), pp.2624–33.

Meyer, K.B. et al., 2008. Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer. *PLoS Biology*, 6(5), p.e108.

Millstein, J. et al., 2009. Disentangling molecular relationships with a causal inference test. *BMC Genetics*, 10(1), p.23.

Minami, S. et al., 2010. Proteomic study of sera from patients with bladder cancer: usefulness of S100A8 and S100A9 proteins. *Cancer Genomics & Proteomics*, 7(4), pp.181–9.

Moffatt, M.F. et al., 2010. A large-scale, consortium-based genomewide association study of asthma. *The New England Journal of Medicine*, 363(13), pp.1211–21.

Morley, M. et al., 2004. Genetic analysis of genome-wide variation in human gene expression. *Nature*, 430(7001), pp.743–7.

Nica, A.C. et al., 2010. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. G. Gibson, ed. *PLoS Genetics*, 6(4).

Nicolae, D.L. et al., 2010. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genetics*, 6(4), p.e1000888.

Van Oers, J.M.M. et al., 2007. FGFR3 mutations and a normal CK20 staining pattern define low-grade noninvasive urothelial bladder tumours. *European Urology*, 52(3), pp.760–8.

Olsson, A.H. et al., 2014. Genome-wide associations between genetic and epigenetic variation influence mRNA expression and insulin secretion in human pancreatic islets. *PLoS Genetics*, 10(11), p.e1004735.

Palermo, G., Piraino, P. & Zucht, H.D., 2009. Performance of PLS regression coefficients in selecting variables for each response of a multivariate PLS for omics-type data. *Adv Appl Bioinform Chem*, 2, pp.57–70.

Parkhomenko, E., Tritchler, D. & Beyene, J., 2009. Sparse canonical correlation analysis with application to genomic data integration. *Stat Appl Genet Mol Biol*, 8, p.Article 1.

Pennisi, E., 2012. Genomics. ENCODE project writes eulogy for junk DNA. *Science (New York, N.Y.)*, 337(6099), pp.1159, 1161.

Peter H. Westfall & Young, S.S., 1993. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, Wiley-Interscience: New York.

Piazza, R. et al., 2013. Recurrent SETBP1 mutations in atypical chronic myeloid leukemia. *Nature Genetics*, 45(1), pp.18–24.

Pickrell, J.K. et al., 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289), pp.768–72.

Pineda, S., Gomez-Rubio, P., et al., 2015. Framework for the Integration of Genomics, Epigenomics and Transcriptomics in Complex Diseases. *Human Heredity*, 79(3-4), pp.124–36.

Pineda, S. et al., 2014. Genetic variation in the TP53 pathway and bladder cancer risk. a comprehensive analysis. M. Katoh, ed. *PloS One*, 9(5), p.e89952.

Pineda, S., Real, F.X., et al., 2015. Integration analysis of three omics data using penalized regression methods: An application to bladder cancer. *PLoS Genetics(Submitted)*.

Portela, A. & Esteller, M., 2010. Epigenetic modifications and human disease. *Nat Biotechnol*, 28(10), pp.1057–1068.

Qi, W. et al., 2013. Inhibition of inducible heat shock protein-70 (hsp72) enhances bortezomib-induced cell death in human bladder cancer cells. *PloS One*, 8(7), p.e69509.

Rakyan, V.K. et al., 2011. Epigenome-wide association studies for common human diseases. *Nature reviews. Genetics*, 12(8), pp.529–41.

Reinert, T. et al., 2011. Comprehensive genome methylation analysis in bladder cancer: identification and validation of novel methylated genes and application of these as urinary tumor markers. *Clinical cancer research : An Official Journal of the American Association for Cancer Research*, 17(17), pp.5582–92.

Reinert, T. et al., 2012. Diagnosis of bladder cancer recurrence based on urinary levels of EOMES, HOXA9, POU4F2, TWIST1, VIM, and ZNF154 hypermethylation. B. C. Christensen, ed. *PloS One*, 7(10), p.e46297.

Riggs, A.D., 1975. X inactivation, differentiation, and DNA methylation. *Cytogenetics and Cell Genetics*, 14(1), pp.9–25.

Ritchie, M.D. et al., 2015. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, 16(2), pp.85–97.

Robertson, K.D. & Wolffe, A.P., 2000. DNA methylation in health and disease. *Nature reviews. Genetics*, 1(1), pp.11–9.

Rothman, N. et al., 2010. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nature Genetics*, 42(11), pp.978–984.

S Pineda, P Gomez-Rubio, A Picornell, K Bessonov, M Márquez, M Kogevinas, FX Real, K Van Steen, N.M., 2015. Framework for the integration of genomics, epigenomics, and transcriptomics in complex diseases. *Human heredity*.

Samanic, C. et al., 2006. Smoking and bladder cancer in Spain: effects of tobacco type, timing, environmental tobacco smoke, and gender. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 15(7), pp.1348–54.

Samanic, C.M. et al., 2008. Occupation and bladder cancer in a hospital-based case-control study in Spain. *Occupational and Environmental Medicine*, 65(5), pp.347–53.

Seo, J.-H. et al., 2013. Deconvoluting complex tissues for expression quantitative trait locus-based analyses. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1620), p.20120363.

Serizawa, R.R. et al., 2011. Integrated genetic and epigenetic analysis of bladder cancer reveals an additive diagnostic value of FGFR3 mutations and hypermethylation events. *Int J Cancer*, 129(1), pp.78–87.

Sharp, A.J. et al., 2011. DNA methylation profiles of human active and inactive X chromosomes. *Genome Research*, 21(10), pp.1592–600.

Shpak, M. et al., 2014. An eQTL analysis of the human glioblastoma multiforme genome. *Genomics*, 103(4), pp.252–63.

Spearman, C., 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1), pp.72–101.

Stigler, S.M., 1989. Francis Galton’s Account of the Invention of Correlation. *Statistical Science*, 4(2), pp.73–79.

Stranger, B.E. et al., 2007. Population genomics of human gene expression. *Nature Genetics*, 39(10), pp.1217–1224.

De Tayrac, M. et al., 2009. Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. *BMC Genomics*, 10, p.32.

The Cancer Genome Atlas Research Network, 2014. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492), pp.315–22.

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series b (Methodological)*, 58(1), pp.267–288.

Trevor Hastie; Rob Tibshirani; Jerome Friedman, 2001. *The Elements of Statistical Learning*, Springer Series in Statistics.

Tribioli, C. et al., 1992. Methylation and sequence analysis around EagI sites: identification of 28 new CpG islands in XQ24-XQ28. *Nucleic Acids Research*, 20(4), pp.727–33.

Uchida, K. et al., 2014. Investigation of HOXA9 promoter methylation as a biomarker to distinguish oral cancer patients at low risk of neck metastasis. *BMC Cancer*, 14(1), p.353.

Venter, J.C. et al., 2001. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507), pp.1304–51.

Visser-Grieve, S., Hao, Y. & Yang, X., 2012. Human homolog of Drosophila expanded, hEx, functions as a putative tumor suppressor in human cancer cell lines independently of the Hippo pathway. *Oncogene*, 31(9), pp.1189–95.

Wagner, J.R. et al., 2014. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biology*, 15(2), p.R37.

Waldmann, P. et al., 2013. Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in Genetics*, 4, p.270.

Watson, I.R. et al., 2013. Emerging patterns of somatic mutations in cancer. *Nature Reviews. Genetics*, 14(10), pp.703–18.

Weiss, J.R. et al., 2007. Concordance of pharmacogenetic polymorphisms in tumor and germ line DNA in adult patients with acute myeloid leukemia. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 16(5), pp.1038–41.

Westra, H.J. et al., 2013. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*, 45(10), pp.1238–1243.

Wild, P.J. et al., 2005. Gene expression profiling of progressive papillary noninvasive carcinomas of the urinary bladder. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 11(12), pp.4415–29.

Wold, S. et al., 1984. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3), pp.735–743.

Wu, Q. et al., 2007. DNA methylation profiling of ovarian carcinomas and their in vitro models identifies HOXA9, HOXB5, SCGB3A1, and CRABP1 as novel targets. *Molecular Cancer*, 6, p.45.

Wu, T.T. et al., 2009. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics (Oxford, England)*, 25(6), pp.714–21.

Wu, X.-R., 2009. Biology of urothelial tumorigenesis: insights from genetically engineered mice. *Cancer Metastasis Reviews*, 28(3-4), pp.281–90.

Wu, X.-R., 2005. Urothelial tumorigenesis: a tale of divergent pathways. *Nature reviews. Cancer*, 5(9), pp.713–25.

Yao, R. et al., 2007. The S100 proteins for screening and prognostic grading of bladder cancer. *Histology and Histopathology*, 22(9), pp.1025–32.

You, J.S. & Jones, P.A., 2012. Cancer genetics and epigenetics: two sides of the same coin? *Cancer Cell*, 22(1), pp.9–20.

Zeegers, M., Rijsdijk, F. & Sham, P., 2004. Adjusting for covariates in variance components QTL linkage analysis. *Behavior Genetics*, 34(2), pp.127–33.

Zhang, D. et al., 2010. Genetic control of individual differences in gene-specific methylation in human brain. *American Journal of Human Genet*, 86(3), pp.411–419.

Zhernakova, D. V et al., 2013. DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. V. G. Cheung, ed. *PLoS Genetics*, 9(6), p.e1003594.

Zhou, H. et al., 2010. Association screening of common and rare genetic variants by penalized regression. *Bioinformatics (Oxford, England)*, 26(19), pp.2375–82.

Zhu, Y., Qiu, P. & Ji, Y., 2014. TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nature Methods*, 11(6), pp.599–600.

SUPPLEMENTARY MATERIAL

Supplementary Material: PART 3 - Chapter 1

Table S3.1.1: Strong correlations between gene expression and methylation located in the same gene.

Table S3.1.2: cis-eQTLs significant (FDR < 5%)

Table S3.1.3: cis-mQTLs significant (FDR < 5%)

Table S3.1.4: Triple relationship (SNP-CpG-Expression)

Tables S3.1.1-S3.1.4 refer to the accompanying CD

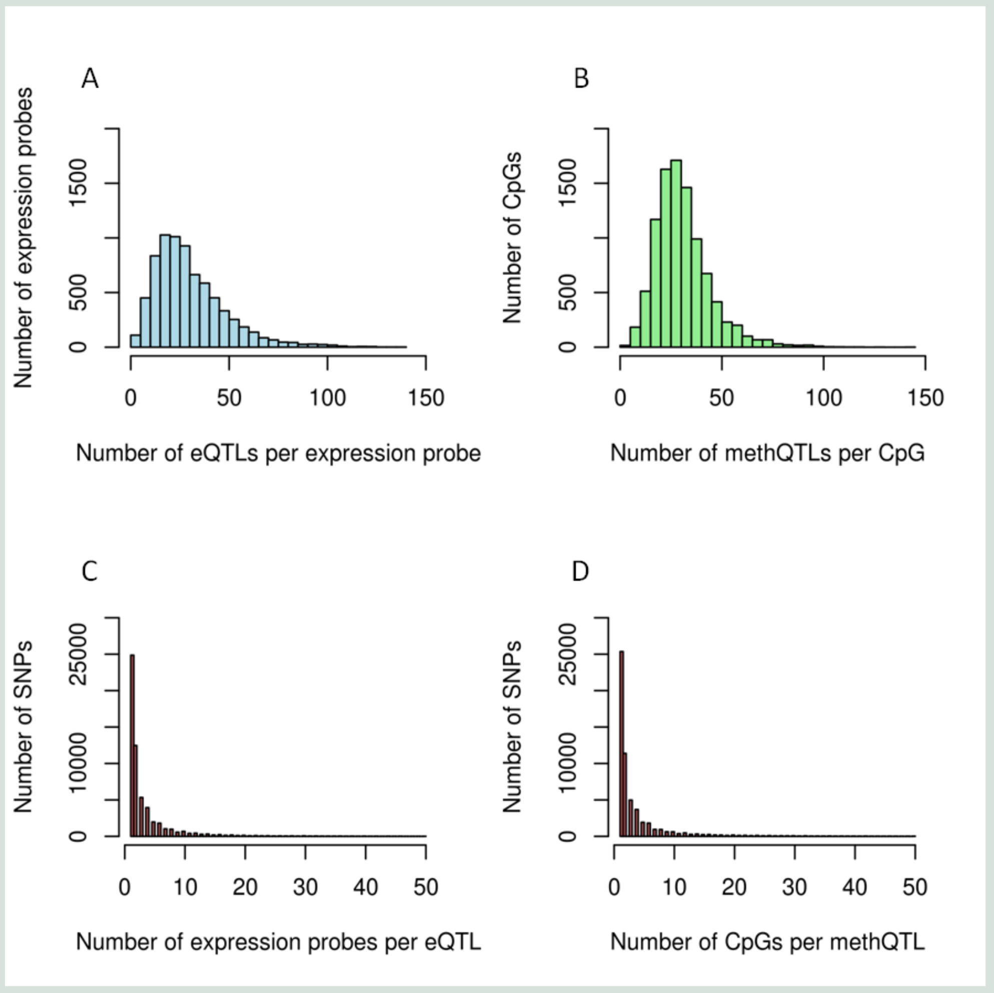


Figure S3.1.1: Distribution of significant multiple QTLs per expression probes (A) and per CpGs (B). Distribution of significant multiple eQTLs (C) and methQTLs (D) per SNPs

Supplementary Material: PART 3 - Chapter 2

S3.2.1 Excel: SNP-LASSO model. Information about SNPs associated with gene expression using LASSO with SNP model for the discovery and replication phase**.**

S3.2.2 Excel: CPG-LASSO model. Information about CpGs associated with gene expression using LASSO with CPG model for the discovery and replication phase**.**

S3.2.3 Excel: Global-LASSO model. Information about SNPs and CpGs associated with gene expression using LASSO with Global model for the discovery and replication phase**.**

S3.2.4 Excel: SNP-ENET model. Information about SNPs associated with gene expression using ENET with SNP model for the discovery and replication phase**.**

S3.2.5 Excel: CPG-ENET model. Information about CpGs associated with gene expression using ENET with CPG model for the discovery and replication phase**.**

S3.2.6 Excel: Global-ENET model. Information about SNPs and CpGs associated with gene expression using ENET with Global model for the discovery and replication phase**.**

S3.2.1-S3.2.6 Excel refer to the accompanying CD

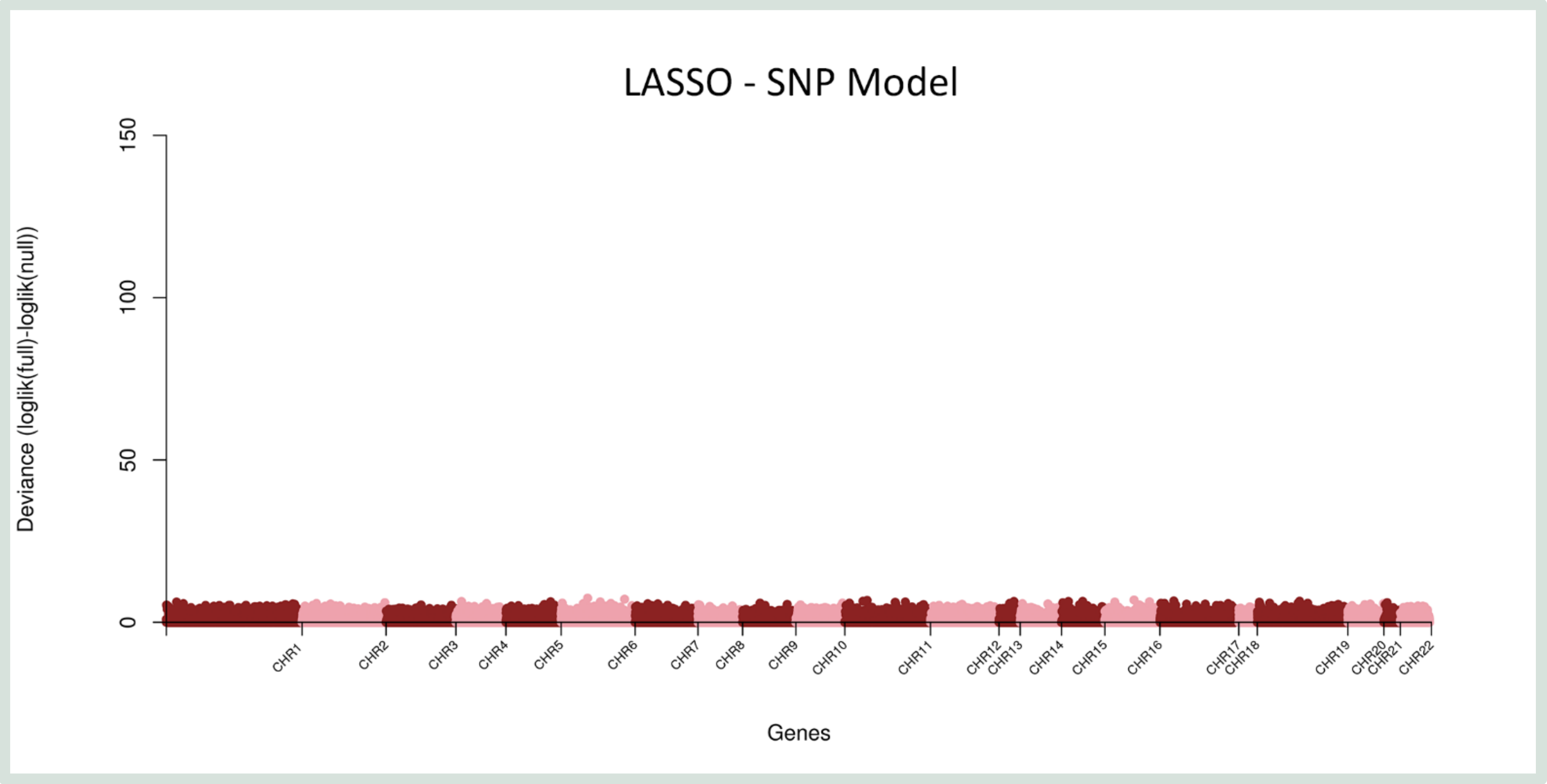


Figure S3.2.1: Deviance across the genome when applying LASSO for the SNP model for the simulated data. The number of genes simulated are 20,899 for 27 individuals using a multivariate normal distribution ($\mu = 8.4$, $\sigma^2 = 0.4$). No gene was significantly associated after the permutation-based MaxT algorithm.

Table S3.2.1: IDs corresponding to the 27 samples from EPICURO data used in this analysis

GSE71666	GSE71576	GSE51641
10090510	10090510	4118698416
10090910	10090910	4235966055
10091310	10091310	4235966030
10091710	10091710	4235966045
10091910	10091910	4118698575
10092810	10092810	4235966022
10093010	10093010	4239166109
10093110	10093110	4235966295
10093210	10093210	4118698560
10093310	10093310	4235966250
10093410	10093410	4118698403
10093510	10093510	4239166219
10093710	10093710	4118698451
10094010	10094010	4235966024
10094310	10094310	4239166062
10094410	10094410	4239166175
30105412	30105412	4118698428
30105711	30105711	4118698441
30106516	30106516	4235966298
30106619	30106619	4118698433
30107012	30107012	4235966076
30107610	30107610	4235966233
30107713	30107713	4235966253
30107919	30107919	4118698426
30108817	30108817	4235966029
30109911	30109911	4118698434
30110214	30110214	4118698427

Table S3.2.2: IDs corresponding to the 238 samples from the TCGA data used in this analysis

A0S7	A0YX	A0F0	A0F6	A0YN	A0YR	A0YO	A1HR
A1A3	A20J	A20N	A20O	A20P	A20Q	A20T	A20U
A0C8	A13J	A1A5	A1A6	A1A7	A1AA	A1AB	A1AC
A1AG	A1AF	A20R	A20X	A2LA	A2LB	A27C	A2HX
A2I6	A2PC	A3B3	A3B4	A3EE	A2EC	A2EF	A2EJ
A2EO	A2ES	A2C5	A2HO	A2HQ	A2OE	A3JX	A3JW
A2LD	A3B6	A3IT	A3IL	A3IN	A3B8	A3MF	A3MI
A3B7	A3MH	A3IB	A3IU	A3IS	A3IM	A2I4	A3IV
A3B5	A3JM	A3JZ	A3N6	A3KJ	A3PH	A3PJ	A3PK
A3OQ	A3OS	A3JV	A3QG	A3QH	A3QI	A3QU	A3YL
A3X1	A3X2	A3X6	A3Y1	A3SJ	A3SL	A3SM	A3SN
A3SQ	A3SR	A3SS	A3VY	A3BM	A3OO	A3RC	A3RD
A3WS	A3WV	A0F1	A0F7	A0EZ	A42C	A3WW	A3ZE
A42R	A40E	A40G	A3Z7	A42F	A42E	A47T	A47S
A47X	A47Y	A43N	A43P	A43S	A43U	A43X	A42P
A5UA	A5W6	A5KE	A5KF	A5BY	A5BZ	A5C0	A5C1
A5RJ	A5Z6	A4IJ	A4XJ	A541	A43Y	A5BR	A5BS
A5BV	A5BX	A3Z9	A4ZW	A2OF	A5ND	A4AC	A54R
A6AV	A6AW	A6B0	A6B1	A6B2	A6B5	A6B6	A4TZ
A677	A678	A62N	A62O	A62P	A62S	A61P	A6I1
A5RH	A6FZ	A6MB	A66R	A6FI	A6FN	A69X	A6DX
A6MF	A7DU	A6TF	A6TG	A6TH	A6TI	A76B	A763
A72E	A7DV	A6TA	A6TB	A6TC	A6TD	A6TE	A6TK
A41N	A41P	A41Q	A41S	A78K	A78L	A78N	A78O
A20V	A1AE	A2I2	A2EL	A3MG	A3NA	A3N5	A3OP
A3SP	A3QF	A3YS	A47W	A5U8	A5RI	A5BU	A5ZZ
A6C6	A6ME	A767	A6ZA	A13I	A20W	A1AD	A2I1
A2EK	A3I6	A3IE	A3IK	A3RB	A3WX	A3SO	A3WC
A3ZF	A47V	A42Q	A5NE	A5BT	A51V	A4U1	A6I3
A3IQ	A766	A762	A1HS	A3WY	A519		

Table S3.2.3: Functional Annotation Clustering from DAVID tool (Enrichment score ≥ 1.3)

Cluster 1	**Enrichment Score: 3.5**				
Category	Term	Count	PValue	Genes	Benjamini
GOTERM_CC_FAT	GO:0005576~extracellular region	17	1.05E-05	*OLFM4, CRTAC1, MSMB, IGJ, MMP7, IGF2, PIGR, TCN1, CXCL17, FREM2, SAA1, REN, IGHD, CRH, PLA2G2A, PTN, CP*	7.79E-04
SP_PIR_KEYWORDS	Secreted	15	1.41E-05	*OLFM4, CRTAC1, MSMB, S100A9, MMP7, IGF2, PIGR, TCN1, CXCL17, SAA1, REN, IGHD, CRH, PTN, CP*	2.02E-03
SP_PIR_KEYWORDS	signal	20	3.81E-05	*OLFM4, CRTAC1, MSMB, IGJ, MMP7, IGF2, PIGR, TCN1, CXCL17, SAA1, FREM2, REN, CRH, CTSE, CEACAM7, PLA2G2A, PTN, CEACAM6, CEACAM5, CP*	2.72E-03
UP_SEQ_FEATURE	signal peptide	20	4.17E-05	*OLFM4, CRTAC1, MSMB, IGJ, MMP7, IGF2, PIGR, TCN1, CXCL17, SAA1, FREM2, REN, CRH, CTSE, CEACAM7, PLA2G2A, PTN, CEACAM6, CEACAM5, CP*	8.55E-03
GOTERM_CC_FAT	GO:0044421~extracellular region part	11	1.02E-04	*OLFM4, CRTAC1, SAA1, FREM2, MSMB, REN, MMP7, PLA2G2A, PTN, IGF2, CP*	3.78E-03
GOTERM_CC_FAT	GO:0005615~extracellular space	9	2.74E-04	*OLFM4, SAA1, MSMB, REN, MMP7, PLA2G2A, PTN, IGF2, CP*	6.74E-03
UP_SEQ_FEATURE	disulfide bond	16	1.08E-03	*OLFM4, CRTAC1, MSMB, IGJ, IGF2, PIGR, TCN1, CXCL17, REN, IGHD, CTSE, CEACAM7, PLA2G2A, PTN, CEACAM6, CP*	1.05E-01

SP_PIR_KEYWORDS	disulfide bond	16	1.48E-03	*OLFM4, CRTAC1, MSMB, IGJ, IGF2, PIGR, TCN1, CXCL17, REN, IGHD, CTSE, CEACAM7, PLA2G2A, PTN, CEACAM6, CP*	6.81E-02
SP_PIR_KEYWORDS	glycoprotein	19	4.63E-03	*SLC38A4, OLFM4, CRTAC1, IGJ, KRT13, IGF2, TSPAN8, PIGR, TCN1, FREM2, REN, IGHD, CTSE, CEACAM7, CEACAM6, CEACAM5, SERPINB4, SERPINB3, CP*	1.24E-01
UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	14	1.37E-01	*SLC38A4, OLFM4, IGJ, TSPAN8, PIGR, TCN1, FREM2, REN, IGHD, CTSE, CEACAM7, CEACAM6, CEACAM5, CP*	9.20E-01
Cluster 2	**Enrichment Score: 1.8**				
Category	Term	Count	PValue	*Genes*	Benjamini
GOTERM_BP_FAT	GO:0032101~regulation of response to external stimulus	4	5.63E-03	*SAA1, PLA2G2A, TSPAN8, IGF2*	7.77E-01
GOTERM_BP_FAT	GO:0050727~regulation of inflammatory response	3	1.30E-02	*SAA1, PLA2G2A, IGF2*	7.52E-01
GOTERM_BP_FAT	GO:0006952~defense response	5	5.03E-02	*SAA1, S100A9, CRH, PLA2G2A, IGF2*	8.40E-01
Cluster 3	**Enrichment Score: 1.7**				
Category	Term	Count	PValue	*Genes*	Benjamini

GOTERM_BP_FAT	GO:0050708~regulation of protein secretion	3	7.76E-03	*SAA1, IGF2, KRT20*	7.49E-01
GOTERM_BP_FAT	GO:0051046~regulation of secretion	4	1.08E-02	*SAA1, CRH, IGF2, KRT20*	7.65E-01
GOTERM_BP_FAT	GO:0060341~regulation of cellular localization	4	1.87E-02	*SAA1, CRH, IGF2, KRT20*	8.13E-01
GOTERM_BP_FAT	GO:0051223~regulation of protein transport	3	2.79E-02	*SAA1, IGF2, KRT20*	7.78E-01
GOTERM_BP_FAT	GO:0070201~regulation of establishment of protein localization	3	3.12E-02	*SAA1, IGF2, KRT20*	7.84E-01
GOTERM_BP_FAT	GO:0032880~regulation of protein localization	3	3.96E-02	*SAA1, IGF2, KRT20*	8.09E-01
Cluster 4	***Enrichment Score: 1.5***				
Category	Term	Count	PValue	*Genes*	Benjamini
GOTERM_BP_FAT	GO:0007610~behavior	6	3.95E-03	*CXCL17, SAA1, REN, S100A9, CRH, PTN*	8.78E-01
GOTERM_BP_FAT	GO:0006935~chemotaxis	3	5.17E-02	*CXCL17, SAA1, S100A9*	8.29E-01

GOTERM_BP_FAT	GO:0042330~taxis	3	5.17E-02	*CXCL17, SAA1, S100A9*	8.29E-01
GOTERM_BP_FAT	GO:0007626~locomotory behavior	3	1.30E-01	*CXCL17, SAA1, S100A9*	9.48E-01
Cluster 5	***Enrichment Score: 1.4***				
Category	Term	Count	PValue	*Genes*	Benjamini
UP_SEQ_FEATURE	region of interest:Coil 2	3	9.54E-03	*KRT5, KRT13, KRT20*	4.82E-01
UP_SEQ_FEATURE	region of interest:Linker 12	3	9.54E-03	*KRT5, KRT13, KRT20*	4.82E-01
UP_SEQ_FEATURE	region of interest:Coil 1B	3	1.13E-02	*KRT5, KRT13, KRT20*	4.44E-01
UP_SEQ_FEATURE	region of interest:Coil 1A	3	1.13E-02	*KRT5, KRT13, KRT20*	4.44E-01
UP_SEQ_FEATURE	region of interest:Linker 1	3	1.13E-02	*KRT5, KRT13, KRT20*	4.44E-01
UP_SEQ_FEATURE	region of interest:Rod	3	1.16E-02	*KRT5, KRT13, KRT20*	3.83E-01
UP_SEQ_FEATURE	region of interest:Head	3	1.23E-02	*KRT5, KRT13, KRT20*	3.46E-01
UP_SEQ_FEATURE	region of interest:Tail	3	1.29E-02	*KRT5, KRT13, KRT20*	3.18E-01
SP_PIR_KEYWORDS	Intermediate filament	3	1.38E-02	*KRT5, KRT13, KRT20*	2.46E-01
INTERPRO	IPR018039:Intermediate filament protein, conserved site	3	1.42E-02	*KRT5, KRT13, KRT20*	2.85E-01

INTERPRO	IPR016044:Filament	3	1.42E-02	*KRT5, KRT13, KRT20*	2.85E-01
INTERPRO	IPR001664:Intermediate filament protein	3	1.46E-02	*KRT5, KRT13, KRT20*	2.41E-01
PIR_SUPERFAMILY	PIRSF002282:cytoskeletal keratin	3	1.78E-02	*KRT5, KRT13, KRT20*	3.84E-01
SP_PIR_KEYWORDS	keratin	3	4.29E-02	*KRT5, KRT13, KRT20*	4.66E-01
GOTERM_CC_FAT	GO:0005882~intermediate filament	3	8.06E-02	*KRT5, KRT13, KRT20*	7.12E-01
GOTERM_CC_FAT	GO:0045111~intermediate filament cytoskeleton	3	8.37E-02	*KRT5, KRT13, KRT20*	6.60E-01
GOTERM_MF_FAT	GO:0005198~structural molecule activity	4	1.79E-01	*KRT5, MYBPC1, KRT13, KRT20*	9.94E-01
GOTERM_CC_FAT	GO:0044430~cytoskeletal part	5	2.29E-01	*TNNT3, KRT5, MYBPC1, KRT13, KRT20*	8.82E-01
GOTERM_CC_FAT	GO:0005856~cytoskeleton	5	4.84E-01	*TNNT3, KRT5, MYBPC1, KRT13, KRT20*	9.93E-01
SP_PIR_KEYWORDS	coiled coil	5	6.92E-01	*OLFM4, KRT5, TRIM31, KRT13, KRT20*	9.99E-01

GOTERM_CC_FAT	GO:0043232~intracellular non-membrane-bounded organelle	6	8.29E-01	*TNNT3, KRT5, MYBPC1, S100A9, KRT13, KRT20*	1.00E+00
GOTERM_CC_FAT	GO:0043228~non-membrane-bounded organelle	6	8.29E-01	*TNNT3, KRT5, MYBPC1, S100A9, KRT13, KRT20*	1.00E+00
Cluster 6	**Enrichment Score: 1.3**				
Category	Term	Count	PValue	*Genes*	Benjamini
GOTERM_BP_FAT	GO:0051046~regulation of secretion	4	1.08E-02	*SAA1, CRH, IGF2, KRT20*	7.65E-01
GOTERM_BP_FAT	GO:0060341~regulation of cellular localization	4	1.87E-02	*SAA1, CRH, IGF2, KRT20*	8.13E-01
GOTERM_BP_FAT	GO:0048585~negative regulation of response to stimulus	3	2.19E-02	*SAA1, CRH, IGF2*	8.14E-01
GOTERM_BP_FAT	GO:0051047~positive regulation of secretion	3	2.57E-02	*SAA1, CRH, IGF2*	7.85E-01
GOTERM_BP_FAT	GO:0006954~inflammatory response	4	3.76E-02	*SAA1, S100A9, CRH, IGF2*	8.17E-01

GOTERM_BP_FAT	GO:0006952~defense response	5	5.03E-02	*SAA1, S100A9, CRH, PLA2G2A, IGF2*	8.40E-01
GOTERM_BP_FAT	GO:0051050~positive regulation of transport	3	9.21E-02	*SAA1, CRH, IGF2*	9.24E-01
GOTERM_BP_FAT	GO:0051240~positive regulation of multicellular organismal process	3	1.07E-01	*SAA1, CRH, IGF2*	9.27E-01
GOTERM_BP_FAT	GO:0009611~response to wounding	4	1.20E-01	*SAA1, S100A9, CRH, IGF2*	9.41E-01
GOTERM_BP_FAT	GO:0007267~cell-cell signaling	4	1.57E-01	*S100A9, CRH, CEACAM6, IGF2*	9.61E-01
Cluster 7	***Enrichment Score: 1.3***				
Category	Term	Count	PValue	*Genes*	Benjamini
GOTERM_MF_FAT	GO:0005509~calcium ion binding	8	4.19E-03	*CAPNS2, ANXA10, CRTAC1, FREM2, S100A9, MMP7, PLA2G2A, S100A2*	3.57E-01
SP_PIR_KEYWORDS	calcium	6	3.55E-02	*CAPNS2, FREM2, S100A9, MMP7, PLA2G2A, S100A2*	4.37E-01
UP_SEQ_FEATURE	domain:EF-hand 1	3	6.58E-02	*CAPNS2, S100A9, S100A2*	7.89E-01
UP_SEQ_FEATURE	domain:EF-hand 2	3	6.58E-02	*CAPNS2, S100A9, S100A2*	7.89E-01

INTERPRO	IPR018249:EF-HAND 2	3	1.02E-01	*CAPNS2, S100A9, S100A2*	6.75E-01
INTERPRO	IPR018247:EF-HAND 1	3	1.04E-01	*CAPNS2, S100A9, S100A2*	6.43E-01
INTERPRO	IPR011992:EF-Hand type	3	1.19E-01	*CAPNS2, S100A9, S100A2*	6.62E-01

Table S3.2.4: Comparison of the deviance, p-value and SNPs and/or CpGs detected by each model between LASSO and ENET methods

		LASSO			ENET		
	GENE	Dev.	p.value	markers detected	Dev.	P.value	markers detected
SNP model	AIM2	55.8	0.1	6	91.8	0.13	18
	CRTAC1	66.2	0.03	18	72.8	0.24	23
	SCNN1A	57.9	0.08	26	54.1	0.55	47
	CLIC6	75.3	0.01	30	75.3	0.17	104
	GSTT1	40.4	0.07	16	43.8	0.9	24
	ANXA10	0	-	-	137.0	0.01	17
	MSMB	4.0	1	3	91.8	0.07	78
	MMP7	0	-	-	94.8	0.06	19
	TCN1	16.3	0.88	1	88.9	0.07	122
	IGF2	10.5	0.98	1	101.6	0.05	55
	GTSF1	50.4	0.23	2	109.6	0.05	19
	IGLJ3	0	-	-	97.7	0.05	182
CPG model	S100A9	52.5	0.08	10	74.6	0.53	42
	S100A2	58.7	0.04	27	58.7	0.66	56
	CRTAC1	60.9	0.1	12	62.3	0.55	12
	SAA1	127.8	0.04	21	102.7	0.31	35
	MYBPC1	74.5	0.08	26	74.5	0.55	34
	SLC38A4	51.7	0.08	16	56.0	0.74	21
	GTSF1	46.7	0.1	3	82.2	0.28	9
	TSPAN8	67.2	0.05	9	69.3	0.55	9

	FREM2	70.2	0.06	16	76.0	0.52	27
	C15orf48	83.7	0.05	25	42.7	0.23	9
	CAPNS2	54.9	0.07	9	66.5	0.48	21
	KRT20	93.7	<0.01	26	93.7	0.11	53
	SERPINB4	68.5	0.03	4	94.0	0.12	18
	CXCL17	46.8	0.1	8	45.5	0.24	7
	CLIC6	75.3	0.07	27	75.3	0.51	31
	TMEM45A	57.3	0.08	13	66.2	0.48	61
	IGJ	59.0	0.09	5	174.6	1	32
	UBD	75.0	0.07	11	75.5	0.51	11
	ARHGEF35	49.6	0.09	9	51.9	0.8	14
	CRH	56.7	0.1	4	60.1	0.59	5
	TRIM31	47.1	0.1	27	40.6	0.32	55
	MMP7	0	-	-	99.4	0.08	64
Global model	S100A9	53.66	0.03	11	46.06	0.59	8
	CTSE	60.7	0.06	12	70.12	0.23	17
	PLA2G2A	71.4	0.01	10	66.78	0.26	32
	HMGCS2	53.3	0.02	8	58.21	0.18	10
	AIM2	61.5	0.06	8	104.88	0.12	24
	PIGR	75.8	<0.01	21	75.48	0.12	21
	TNNT3	44.9	0.09	26	36.19	0.82	59
	SCNN1A	58.8	0.03	29	58.76	0.18	31
	KRT5	58.2	0.02	25	58.14	0.18	31

	FREM2	46.0	0.08	2	48.24	0.45	2
	OLFM4	60.0	0.06	10	61.90	0.16	11
	C15orf48	49.9	0.02	7	48.80	0.41	6
	KRT20	48.4	0.05	3	39.50	0.74	1
	KRT13	53.6	0.02	8	54.01	0.25	8
	CEACAM7	76.0	<0.01	19	77.40	0.16	32
	CLIC6	45.1	0.09	4	47.35	0.50	6
	CP	51.1	0.02	3	47.98	0.47	3
	IGJ	58.4	0.03	2	94.72	0.17	16
	CRABP2	9.78	0.99	2	65.2	0.09	26
	REN	0	-	0	84.3	0.03	22
	IGF2	89.01	0.17	11	92.1	0.04	15
	CEACAM5	90.85	0.19	22	92.1	0.06	26
	NLRP2	14.26	0.93	2	84.2	0.04	34

Table S3.2.5: Comparison of genes selected by each model (SNP, CpG and Global model) using LASSO

	Global model				SNP model			CPG model		
GENE	Dev	P	SNPs	CPGs	Dev	P	SNPs (common)	Dev	P	CPGs (common)
GSTM1	79.9	0.03	12	5	0.0	-	0	0.0	-	0
TMEM45A	57.3	0.01	12	1	53.1	0.23	14 (11)	1.2	1.0	1 (1)
ANXA10	153.3	0.01	22	0	0.0	-	0	0.0	-	0
ALDH7A1	41.1	0.05	6	0	41.1	0.36	6 (6)	6.2	1.0	1 (0)
UBD	53.7	0.04	2	4	3.1	1.0	1 (1)	91.0	0.22	13 (4)
PTN	81.6	0.04	14	0	79.3	0.04	12 (12)	0.0	-	0
IGF2	77.95	0.02	8	2	31.3	0.21	4 (3)	92.3	0.23	13 (2)
SLC38A4	57.8	0.01	18	2	0.0	-	0	13.3	0.94	3 (2)
SERPINB4	78.0	0.02	6	0	91.7	<0.01	13 (6)	17.0	0.90	1 (0)
SERPINB3	142.3	0	1	0	171.6	0.80	29 (11)	25.2	0.46	1 (0)
CEACAM5	88.9	0.02	13	5	0.0	-	0 (0)	77.6	0.05	16 (5)
AIM2	10.7	0.95	1 (1)	0	107.5	0.02	24			
FCGR3A	54.5	0.55	23 (14)	4	45.1	0.04	24			
AGMO	51.4	0.06	18 (18)	0	45.2	0.04	13			

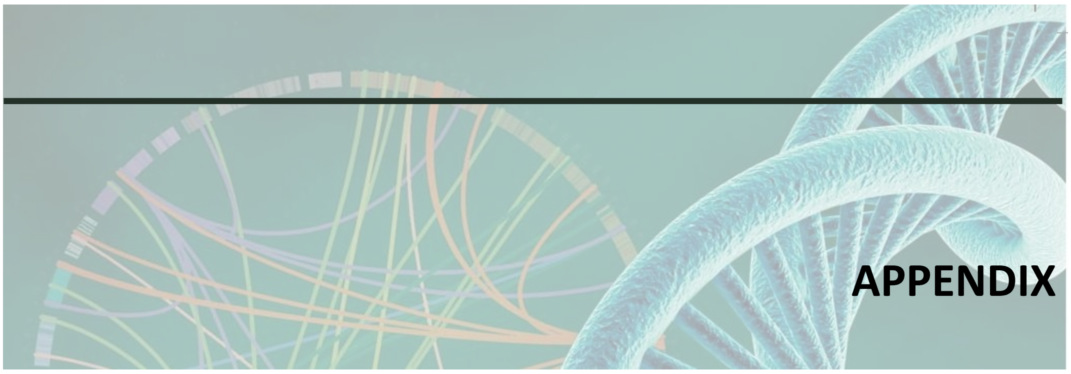
PTN	81.6	0.04	14 (11)	0	79.3	0.04	12			
ARHGEF35	48.4	0.10	7 (5)	1	40.5	0.05	5			
SAA2	16.8	0.72	5 (3)	1	69.3	0.04	13			
IGHD	66.9	0.11	9 (9)	2	71.7	0.01	10			
SERPINB4	78.0	0.02	6 (6)	0	91.7	0	13			
CEACAM6	14.7	1.0	0	1	70.4	0.02	9			
CLIC6	75.3	0.09	25 (14)	2	73.0	0.02	21			
PLA2G2A	88.6	0.77	24	6 (6)				72.9	0.04	12
HMGCS2	0.0	-	0	0				58.7	0.04	10
S100A8	63.1	0.10	3	4 (3)				55.1	0.04	4
AIM2	10.7	0.95	1	0				70.3	0.04	10
PIGR	0.0	-	0	0				65.0	0.04	9
IGJ	59.0	0.16	3	2 (2)				70.8	0.05	4
BHMT	0	-	0	0				49.4	0.05	9
LCN2	70.7	0.08	11	6 (5)				49.7	0.05	6
MSMB	0.0	-	0	0				77.3	0.02	8

TCN1	21.7	0.87	1	1 (1)		55.1	0.04	8
KRT5	12.3	0.98	0	1 (1)		58.2	0.04	25
CAPNS2	62.7	0.08	15	1 (1)		50.7	0.04	7
KRT13	63.3	0.09	8	4 (4)		52.3	0.04	7
C3	42.1	0.19	10	4 (3)		629	0.04	21
CEACAM7	44.7	0.28	2	3 (2)		77.5	0.02	25
CEACAM5	88.9	0.02	13	5 (5)		77.6	0.05	16
NLRP2	0.0	-	0	0		73.1	0.05	16

Supplementary Material: PART 3 - Chapter 3

S3.3.1 Excel: Genes and associated eQTLs from model 1, 2 and 3.

S3.3.1 Excel refer to the accompanying CD



APPENDIX

Published papers during my PhD

Silvia Pineda, Paulina Gomez-Rubio, Antoni Picornell, Kyrilo Bessonov, Mirari Márquez, Manolis Kogevinas, Francisco X Real, Kristel Van Steen, Núria Malats. "Framework for the integration of genomics, epigenomics, and transcriptomics for complex diseases" Hum Hered 2015;79:124-136. DOI:10.1159/000381184

Silvia Pineda, Roger L. Milne, M. Luz Calle, Nathaniel Rothman, Evangelina López de Maturana, Jesús Herranz, Manolis Kogevinas, Stephen J. Chanock, Adonina Tardón, Mirari Márquez, Lin T. Guey, Montserrat García-Closas, Josep Lloreta, Erin Baum, Anna González-Neira, Alfredo Carrato, Arcadi Navarro, Debra T. Silverman, Francisco X. Real Núria Malats. "Genetic variation in the TP53 pathway and bladder cancer risk. A comprehensive analysis" PLoS One. 2014 May 12; 9(5):e89952. DOI:10.1371/journal.pone.0089952

Divyansh Agarwal, **Silvia Pineda**, et al. "FGF receptor genes and breast cancer susceptibility: results from the Breast Cancer Association Consortium", British Journal of Cancer (February 2014), 1–13 | DOI: 10.1038/bjc.2013.769

Submitted papers

Silvia Pineda, Francisxo X Real, Manolis Kogevinas, Alfredo Carrato, Stephen J. Chanock, Núria Malats, Kristel Van Steen. "Integration analysis of three *omics* data using penalized regression methods: An application to bladder cancer" PlosGenetics, 2015

Evangelina López de Maturana, **Silvia Pineda,** Angela Brand, and Núria Malats. Towards the integration of omics data in epidemiological studies: still a "long and winding road." Genetic Epidemiology, (2015)

Silvia Pineda, Kristel Van Steen, Núria Malats. "Integrative eQTL –omics analysis considering tumor tissue and blood samples in individuals with bladder cancer" AJHG, 2015

Participation in conference during my PhD

Capita Selecta in Complex Disease Analysis **Nov. 24-26, 2014**

Liège, Belgium

Oral Presentation: Integration analysis of ‘omics’ data with penalized regression: An application in bladder cancer

SEE: XXXII Scientific Meeting of the Spanish Epidemiology Society **Sept. 3-5, 2014**

Alicante, Spain

Oral Presentation: Integration analysis of ‘omics’ data with penalized regression: An application in bladder cancer

EMGM: European Mathematical Genetics Meeting **April 1-2, 2014**

Köln, Germany

Poster: Integration analysis of ‘OMICS’ data using penalized regression methods: An application to bladder cancer

XIV Scientific Meeting of the Spanish Biometrics Society **May 22-24, 2013**

Ciudad Real, Spain

Oral presentation: Statistical approaches for the integration of ‘omics’ and epidemiological data: an application to bladder cancer.

SEE: XXX Scientific Meeting of the Spanish Epidemiology Society **Oct. 17-19, 2012**

Santander, Spain

Oral presentation: The application of penalized regression (LASSO) for genetic association studies

5TH International Consortium of Bladder Cancer Meeting **Oct. 8-9, 2012**

Spanish National Cancer Research Centre (CNIO), Madrid, Spain

Poster: Eped+omics, statistical approaches for the integration of ‘omics’ data: An application in bladder cancer

Invited talks & seminars during my PhD

Cancer Genomics and Personalized Medicine Workshop **July 1, 2015**

Barcelona, Spain

Invited Speaker: Adapting innovative statistical modelling for integrative –omics analysis towards the identification of cancer molecular insights

Statistical Modeling of Cancer Genetic Predisposition

Vienna, Austria **Mar. 19-20, 2015**

Invited Speaker: Approaching the complexity of bladder cancer genetic predisposition by applying innovative statistical modelling

Teaching during my PhD

Master Degree in Molecular Oncology **April 25, 2015**

Bio-health studies center, Madrid, Spain.

Course: Survival Analysis

Mentor of students from medical degree in the preventive medicine and public health course **April, 2015**

Autónoma University, Madrid, Spain.

Master in Bioinformatics and computational biology **Feb. 3-4, 2015**

Institute of Health Carlos III (ISCIII), Madrid, Spain

Course: Genomic Variation

Original Paper

Human Heredity

Hum Hered 2015;79:124–136
DOI: 10.1159/000381184

Published online: July 28, 2015

Framework for the Integration of Genomics, Epigenomics and Transcriptomics in Complex Diseases

Silvia Pineda[a, f] Paulina Gomez-Rubio[a] Antonio Picornell[a] Kyrylo Bessonov[f] Mirari Márquez[a] Manolis Kogevinas[c, d] Francisco X. Real[b, e] Kristel Van Steen[f, g] Nuria Malats[a]

[a]Genetic and Molecular Epidemiology Group, and [b]Epithelial Carcinogenesis Group, Spanish National Cancer Research Centre (CNIO), Madrid, [c]Centre for Research in Environmental Epidemiology (CREAL), [d]Institut Municipal d'Investigació Mèdica – Hospital del Mar, and [e]Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain; [f]Systems and Modeling Unit, Montefiore Institute, University of Liège, and [g]Bioinformatics and Modeling, GIGA-R, University of Liège, Liège, Belgium

Key Words

Bladder cancer · 'Omics' · Integration · Genomics · Epigenomics · Transcriptomics

Abstract

Objectives: Different types of 'omics' data are becoming available in the post-genome era; still a single omics assessment provides limited insights to understand the biological mechanism of complex diseases. Genomics, epigenomics and transcriptomics data provide insight into the molecular dysregulation of neoplastic diseases, among them urothelial bladder cancer (UBC). Here, we propose a detailed analytical framework necessary to achieve an adequate integration of the three sets of omics data to ultimately identify previously hidden genetic mechanisms in UBC. ***Methods:*** We built a multi-staged framework to study possible pair-wise combinations and integrated the data in three-way relationships. SNP genotypes, CpG methylation levels and gene expression levels were determined for a total of 70 individuals with UBC and with fresh tumour tissue available. ***Results:*** We suggest two main hypothesis-based scenarios for gene regulation based on the omics integration analysis, where DNA methylation affects gene expression and genetic variants co-regulate gene expression and DNA methylation. We identified several three-way trans-association 'hotspots' that are found at the molecular level and that deserve further studies. ***Conclusions:*** The proposed integrative framework allowed us to identify relationships at the whole-genome level providing some new biological insights and highlighting the importance of integrating omics data.

Introduction

Many data in the molecular field ('omics' data) are being generated at an unprecedented pace, this including genome, methylome, transcriptome, and microbiome, among others. There is a growing interest in combining

KARGER 125 YEARS

E-Mail karger@karger.com
www.karger.com/hhe

© 2015 S. Karger AG, Basel
0001–5652/15/0794–0124$39.50/0

Nuria Malats, MD, MPH, PhD
Genetic and Molecular Epidemiology Group
Spanish National Cancer Research Centre (CNIO)
C/Melchor Fernández Almagro, 3, ES–28029 Madrid (Spain)
E-Mail nmalats@cnio.es

Prof. Kristel Van Steen, PhD, PhD
Systems and Modeling Unit, Montefiore Institute
University of Liège, Bât. B28 Bioinformatique, Grande Traverse 10
BE–4000 Liège (Belgium)
E-Mail kristel.vansteen@ulg.ac.be

the different types of omics data sets that are becoming available, since a single omics assessment provides limited insights into the understanding of the underlying biological mechanisms of a physiological/pathological condition. For example, even when many genome-wide association studies (GWAS) have identified several SNPs involved in complex diseases, the functional implications of the susceptibility loci are still poorly understood, and they only partially account for the phenotype variability. Combining different omics data types seems to be a more suitable approach, as it will likely reveal previously hidden information.

The simplest form of data integration involves the combination of two different data types, common examples being genetic variants and gene expression or, more recently, genetic variants and DNA methylation [1]. DNA methylation involves the addition of a methyl group to the 5′ position of the cytosine at a CpG site. Genomic regions with a high density of CpG dinucleotides are denominated CpG islands; they are often located in gene promoters and have important roles in gene regulation. CpG sites located up to 2 kb from the island's boundaries are called CpG shores, and it has been demonstrated that they are also very important for gene regulation and that they are implicated in cancer [2]. Both CpG islands and shores, when hypermethylated and located in the promoter region of a gene, negatively regulate gene repression [3]. Therefore, it is important to take the relationship between DNA methylation and gene regulation into account in order to better understand complex diseases [4]. For example, it has been shown that hypermethylation of CpGs located in the promoter region of some tumour suppressor genes (*INK4A*, *Rb*, *VHL*, *hMLH1*, or *BRCA1*, etc.) contribute to cancer development [5]. Therefore, analyzing gene expression data without considering epigenetics provides an incomplete genomic explanation of the transcriptome. Moreover, as DNA methylation regulates gene expression, genetic variants affecting CpG sites might, in turn, affect gene expression, too. It is well known that genetic variants can alter gene expression levels and hence the importance of connecting the DNA sequence to the RNA level. The identification of these expression quantitative trait locus (eQTL) relationships may help to identify regulators of gene expression [6]. These eQTLs have been extensively studied to find associations between common genetic variants and gene expression levels [7–11]. By contrast, the study of potential associations between common variants, DNA methylation levels (methylation QTLs or methQTLs) and gene expression has generated less interest so far [1, 12–15].

Genome, transcriptome, and methylome data offer unique opportunities when combined in the same analyses. This strategy has been applied to HapMap cell lines [14], whole blood from healthy human subjects [16] and human monocytes [17]. Furthermore, some studies have combined these types of data to better understand complex diseases such as breast cancer [18] or type 2 diabetes [19]. As DNA methylation is tissue-specific, these analyses have also been applied to different types of tissues such as the human brain [12] or adipose tissue [15]. It is worth noting that the majority of these studies have only assessed *cis*-relationships, but *trans*-effects deserve further study within the omics context, especially as the complex organization of chromatin in the nucleus is better understood.

In the present study, we built and propose a multi-staged analytical framework to integrate omics data. We tested it on an urothelial bladder cancer (UBC) model using common genetic variants, DNA methylation and gene expression transcripts data from 70 cancer patients. We prove the ability of the framework to identify some multi-omics relationships that provide further knowledge to better understand the biological mechanisms underlying the disease.

Material and Methods

Study Subjects

SNP genotypes, CpG methylation levels and gene expression levels were measured for a total of 70 individuals with fresh tumour tissue available who were recruited as part of the pilot phase of the EPICURO study. All of them were histologically confirmed UBC cases recruited at 2 hospitals in Spain during 1997–1998. Tumour DNA and RNA were extracted and used for omics assessment. SNP data were available for 46 patients, CpG methylation for 46 patients and gene expression for 43. The overlapping of patients between the three omics was 31 for the expression-methylation relationship, 27 for the eQTL and 46 for the methQTL studies.

SNP Genotype Data

Genotyping was performed using Illumina HumanHap 1M array in tumour samples. A total of 1,047,101 SNPs were genotyped in 46 individuals. For genotype calling, we used the cluster file obtained when the same array was applied to germline DNA from 2,424 subjects included in the main EPICURO study. We considered SNPs with <5% of missing values and with a minor allele frequency (MAF) of ≥0.01. Standard quality control (QC) was performed using BeadStudio and R. From BeadStudio, the genotypes AA, Aa and aa were obtained in forward strand for those samples having a call rate of >90%.

DNA Methylation Data

After bisulphite modification of 46 tumour DNA samples using EZ-96 DNA Methylation-Gold kit (Zymo Research, Irvin, Calif., USA), CpG methylation data were generated using the Infinium Human Methylation 27 BeadChip Kit that detected the CpG sites with two probes, one designed against the unmethylated site (signal U) and the other against the methylated site (signal M). The level of methylation was determined at each locus by the intensity of the two possible fluorescent signals [20]. At each CpG site, the methylation levels were measured with the β value, defined as:

$$\beta = \frac{\max\left(M, 0\right)}{\max\left(U, 0\right) + \max\left(M, 0\right) + 100}.$$

The maximum between signal intensity and 0 is used for the calculation of β to avoid the negative numbers caused by background subtractions. Consequently, the β values rank between 0 (unmethylated) and 1 (methylated). The constant 100 was used to regularize the β values when they were very small. Although the β values are useful under some circumstances, it has been demonstrated that the M values are statistically more valid than the β values due to a better approximation of the homocedasticity [21]. This property is important when applying regression models that require this assumption. The M value is calculated as follows:

$$M = \log_2\left(\frac{\max\left(M, 0\right) + 1}{\max\left(U, 0\right) + 1}\right).$$

It ranges between $-\infty$ (unmethylated) and $+\infty$ (methylated). In our study, the M values were used when applying linear regression models, while the β values were used in the rest of the analyses.

The initial number of CpGs in the studied array was 27,578. We then applied BeadStudio software and R to pre-process the data. Background normalization was performed minimizing the amount of variation in the background signals between arrays and, as recommended by Illumina, CpGs were rejected when their detection p value was >0.05. β values <0 or >1 were also excluded. CpGs with SNPs (n = 908) or cross-reactive probes (n = 2,985) were deleted based on earlier reports for the 27K array [22]. After QC, a total of 23,034 CpGs were kept for analysis. These were classified into 3 categories for subsequent analyses: CpG islands (located in the promoter region of a gene), CpG shores (in a sequence up to 2 kb from an island) and CpGs outside of an island or a shore.

Gene Expression Data

Gene expression data were obtained from 43 tumour samples using the Affymetrix DNA Microarray Human Gene 1.0 ST Array with 32,321 probes. This array was based on the 2006 (UCSC hg18, NCBI build 36) human genome sequence with coverage of RefSeq, Ensembl and putative complete CDS GenBank transcripts (www.affymetrix.com). QC was performed using Bioconductor libraries in R (www.bioconductor.org/). The arrayQualityMetrics package [23] was used to implement a background correction and to carry out normalization of expression levels across arrays. The application of QC steps resulted in 20,899 probes and 37 individuals. The affy library in R [24] was used to annotate the probes.

Statistical Analysis

First, tumoural DNA methylation levels in CpG sites and gene expression levels were compared using Spearman's rank correlation for non-normally distributed variables. Second, we assessed the eQTLs and methQTLs, via linear regression modelling for those expression-methylation pair probes that were strongly associated in the previous step. To perform these analyses, we obtained a linear regression model for each SNP as:

$$\mathit{Gene\ expression}_i = \alpha + \beta \times \mathit{SNP}_i$$
$$\mathit{Methylation\ CpG}_i = \alpha + \gamma \times \mathit{SNP}_i.$$

Prior to analysis, we excluded those SNPs that had <2 individuals per genotype due to the imbalance that may produce a highly differential gene expression values, i.e. an individual with rare homozygous genotype and with an extreme gene expression value could produce an artificially highly significant p value.

Expression-methylation probe pairs and eQTLs and methQTLs were classified into three categories according to possible genomic distance effects: *cis*-acting, if the probes were located within 1 Mb; *trans*-acting, if the probes were on the same chromosome but located more than 1 Mb apart, and *trans*-acting-outside, if they were on different chromosomes. To control the analyses for multiple testing, we applied Benjamini and Yekutieli's [25] FDR method that allows for panel dependencies between tests. We applied this correction, taking the number of tests performed in the eQTL and the methQTL study independently into account. Finally, we checked the regions of the trait-associated SNPs already published for UBC.

Third, in line with the study, we integrated the results obtained from pair-wise analyses on genome, epigenome and trascriptome data. We checked the SNPs that were common in the eQTL and methQTL analysis based on those probes-CpGs that were previously correlated in order to have a complete view of the genome in individuals with UBC. We obtained the distribution of the triplets (SNP-CpG-gene expression) that were significantly associated in the same relationship.

Statistical analyses were performed with R, and the results were visualized with Circos software [26].

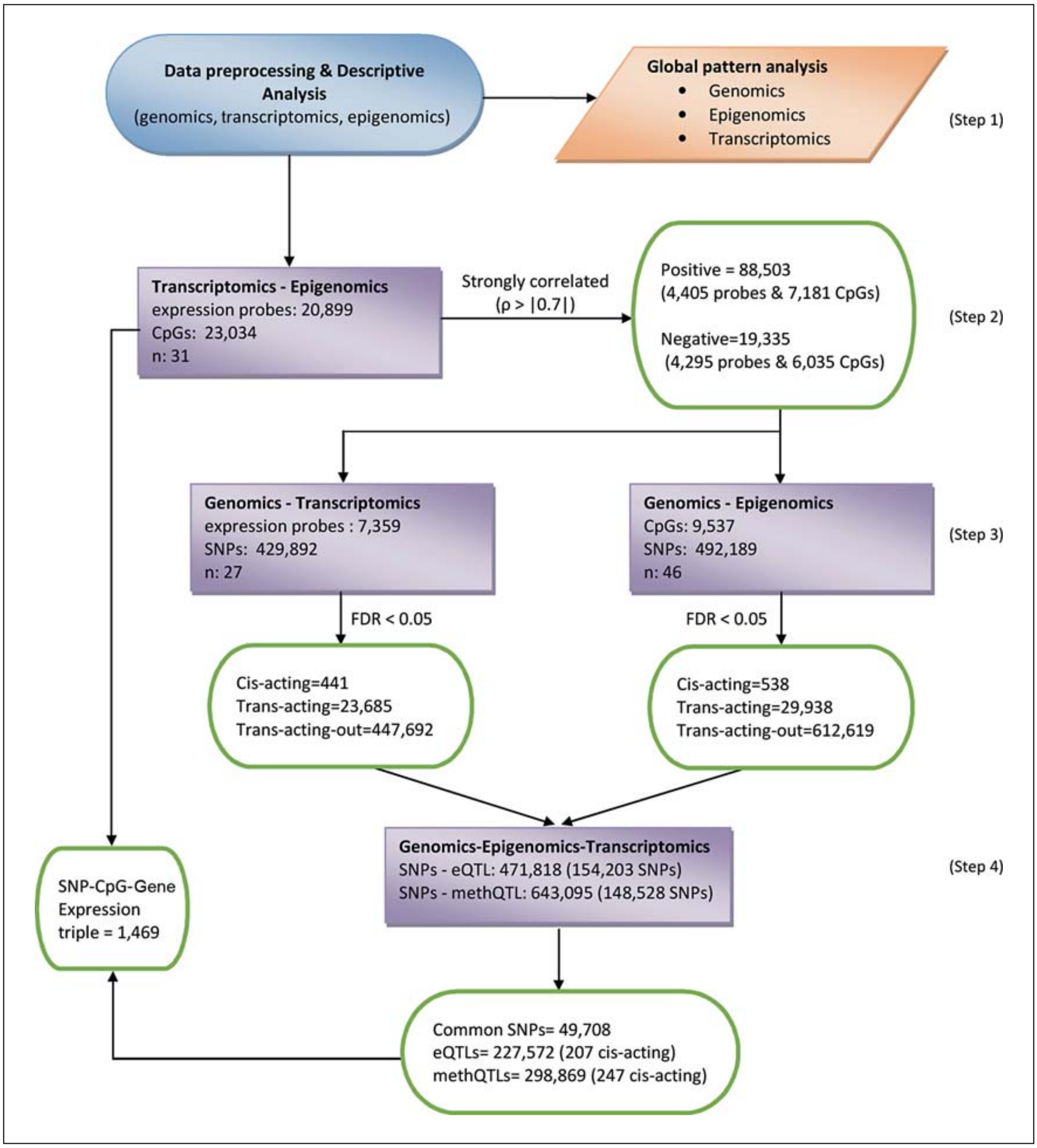
Results

The majority of the individuals included in our study were male (93%) and current (50%) or former (36%) smokers. According to established criteria based on tumour stage and grade for UBC, the individuals were classified as having low-risk non-muscle-invasive tumours (45%), high-risk non-muscle-invasive tumours (22%) or muscle-invasive tumours (29%) (table 1).

The description of the study results is organized in four sections following the framework steps proposed (fig. 1): (1) description of the patterns of individual omics data, globally and according to epidemiological data; (2) correlation analysis between methylation and expression probes; (3) identification of *cis*- and *trans*-eQTLs and methQTLs, and (4) integration of results derived from the previous pair-wise analysis.

Patterns of Individual Omics Data

Table 2 shows the distribution of the genotypes according to their MAF; 14% had a MAF of 0 and were excluded



Color version available online

Fig. 1. Framework for data integration showing the steps to integrate genetic variants, DNA methylation levels and gene expression levels. Step 1 corresponds to the pre-processed data, QC and global patterns individually per data set. Steps 2, 3 and 4 include square boxes corresponding to the analyses performed and the input data and oval boxes corresponding to the results and the input for the next step.

from the analysis, 11% had a MAF of (0.01–0.05], 30% had a MAF of (0.05–0.2], and 31% had a MAF of (0.2–0.4]. Missing values of <5% were observed in 84% of the SNPs.

The patterns for DNA methylation according to the β and M values were different for autosomal chromosomes and X-chromosomes in females due to the X-chromosome inactivation in females. The majority (71%) of CpGs in autosomal chromosomes were unmethylated ($\beta < 0.3$), while, as expected, the majority of the CpGs (66%) in the X-chromosomes showed β values in the range of $0.3 \leq \beta < 0.7$. While the M values for autosomal chromosomes displayed a bimodal distribution, the X-chromosomes ap-

Table 1. Characteristics of the studied patients

Total	72
Gender	
Male	67 (93%)
Female	5 (7%)
Age, years	
Mean ± SD	65.6±9.5
Min.–max.	41–80
Region	
Barcelona	31 (43%)
Elche	41 (57%)
Smoking status	
Non-smokers	8 (11%)
Current smokers	36 (50%)
Former smokers	26 (36%)
Unknown	2 (3%)
Tumor stage	
Low-grade non-muscle-invasive UBC	32 (45%)
High-grade non-muscle-invasive UBC	16 (22%)
Muscle-invasive UBC	21 (29%)
Unknown	3 (4%)

Table 2. Summary of SNPs genotyped

Total number	1,047,101
MAF	
[0.0]	150,548 (14%)
(0.0–0.01]	0 (0)
(0.01–0.05]	108,496 (11%)
(0.05–0.2]	312,220 (30%)
(0.2–0.4]	327,762 (31%)
(0.4–1.0]	148,075 (14%)
Missing values	
No values missing	488,288 (47%)
5% missing values	400,918 (38%)
20% missing values	147,732 (14%)
>20% missing values	10,163 (1%)

MAF = [0.0] means that the individual is common homozygous for the measured SNP.

proximated a normal distribution (online suppl. fig. 1; see www.karger.com/doi/10.1159/000381184 for all online suppl. material). No significant different methylation patterns were found according to the clinical/epidemiological data considered, i.e. smoking status, tumour stage, age, and sex (Pearson's χ^2 test, data not shown).

The expression of the gene probes after background correction and normalization followed a normal distribution (online suppl. fig. 2). We did not find any significant difference according to the clinical/epidemiological data by applying Student's t test (data not shown).

Table 3. Strength of correlations between gene expression and DNA methylation

Spearman's rho	Strength of correlation	Combinations, n
(–0.9 to –1.0]	very strong negative	0
(–0.7 to –0.9]	strong negative	19,335
(–0.4 to –0.7]	moderate negative	9,266,544
(–0.0 to –0.4]	weak negative	238,601,864
[0.0]	no correlation	380,834
(0.0 to 0.4]	weak positive	223,165,638
(0.4 to 0.7]	moderate positive	9,864,848
(0.7 to 0.9]	strong positive	88,503
(0.9 to 1.0]	very strong positive	0

Table 4. Strong correlation for *cis*-acting and *trans*-relationships between CpG methylation and gene expression

		Correlation, n (%)	
		negative	positive
Cis-acting (same gene)	CpG island/shore	37 (80)	9 (20)
	CpG outside	3 (37)	5 (63)
Cis-acting (different gene)	CpG island/shore	41 (26)	116 (74)
	CpG outside	11 (21)	41 (79)
Trans-acting	CpG island/shoe	757 (17)	3,736 (83)
	CpG outside	412 (24)	1,272 (76)
Trans-acting-outside chromosome	CpG island/shore	11,860 (16)	63,054 (84)
	CpG outside	6,214 (23)	20,270 (76)

Correlation between Gene Expression and DNA Methylation

While it is well established that DNA methylation may affect the expression of a gene, mainly in a *cis*-relationship, little is known about *trans*-relationships. We investigated a total of 481,387,566 possible correlations between gene expression and methylation both in *cis*- and in *trans*-relationships. The number of comparisons performed was based on data derived from 31 individuals (table 3). We obtained 19,335 strong negative ($\rho < -0.7$) and 88,503 strong positive ($\rho > 0.7$) associations between gene expression and methylation, corresponding to 7,359 expression traits and 9,537 CpG sites. The distribution of the stronger relationships according to the CpG location and direction is shown in table 4: 5,414 (28%) were located in CpG islands, 1,690 (59%) in CpG shores and 2,433 (57%) outside of CpG islands or shores. There were

Table 5. Significant (FDR <0.05) *cis*-eQTLs and *trans*-eQTLs by MAF and sign of the association

MAF	Sign	*cis*-eQTL, n (%)	*trans*-eQTL, n (%)	*trans-out*-eQTL, n (%)
(0.01–0.2]	Positive	106 (0.005)	7,026 (0.005)	127,177 (0.004)
	Negative	56 (0.002)	2,857 (0.002)	61,134 (0.002)
(0.2–0.4]	Positive	95 (0.003)	4,759 (0.003)	88,213 (0.003)
	Negative	66 (0.002)	3,220 (0.002)	65,457 (0.002)
>0.4	Positive	57 (0.003)	2,930 (0.002)	54,087 (0.002)
	Negative	61 (0.003)	2,893 (0.002)	51,624 (0.002)

Values in parentheses are percentages of significant eQTLs after multiple testing correction over the total number of *cis-* (2,331,808), *trans-* (151,738,928) and *trans-out-* (3,009,504,492) eQTLs.

Table 6. Significant (FDR <0.05) *cis*-methQTLs and *trans*-methQTLs by MAF and sign of the association

MAF	Sign	*cis*-methQTL, n (%)	*trans*-methQTL, n (%)	*trans-out*-methQTL, n (%)
(0.01–0.2]	Positive	137 (0.004)	8,576 (0.004)	190,221 (0.004)
	Negative	61 (0.002)	3,554 (0.002)	72,611 (0.002)
(0.2–0.4]	Positive	118 (0.003)	6,864 (0.003)	139,830 (0.003)
	Negative	139 (0.004)	5,230 (0.002)	98,068 (0.002)
>0.4	Positive	39 (0.001)	3,090 (0.001)	57,476 (0.001)
	Negative	44 (0.001)	2,624 (0.001)	54,413 (0.001)

Values in parentheses are percentages of significant methQTLs after multiple testing correction over the total number of *cis-* (3,499,636), *trans-* (224,328,090) and *trans-out-* (4,466,178,767) methQTLs.

263 (0.03%) *cis*-acting correlations, 6,177 (0.02%) *trans*-acting correlations within the same chromosome and 101,398 (0.02%) *trans*-acting outside the chromosome (*trans-out* correlations). A whole list of CpGs with significant *cis*- association with a gene can be found in online supplementary table 1.

Identification of cis- and trans-eQTLs and methQTLs

In order to detect genetic variants affecting gene expression or DNA methylation, we investigated a total of 7,359 expression traits and 9,537 CpG sites that were strongly correlated in the previous step. The number of SNPs considered here after QC was 429,892 for the eQTL and 492,189 for the methQTL analyses, resulting in a total of 3,163,575,228 eQTLs in 27 individuals and 4,694,006,493 methQTLs explored in 46 individuals. After correction for multiple testing (FDR <0.05), we obtained 471,818 significant eQTLs involving 154,203 SNPs, and 643,095 methQTLs involving 148,528 SNPs. These results point to the fact that multiple expression probes and CpGs were significantly associated with more than one SNP. We refer to this phenomenon as 'hotspots' (online suppl. fig. 3).

We show the distribution of QTLs classified by genomic distance and MAF of the relationship for eQTLs in table 5 and for methQTLs in table 6. When classifying the QTLs by genomic distance, we observed 441 *cis*-eQTLs (0.02%), 23,685 *trans*-eQTLs (0.01%) and 447,692 *trans-out*-eQTLs (0.01%); 538 *cis*-methQTLs (0.01%), 29,938 *trans*-methQTLs (0.01%), and 612,619 *trans-out*-methQTLs (0.01%). When classifying the QTLs in terms of MAF, the majority had a MAF of ≤0.2 (0.006%), while 0.003 and 0.002% had MAFs of (0.2–0.4] and ≥0.4, respectively. Detailed information regarding the *cis*-relationship is provided in online supplementary tables 2 and 3.

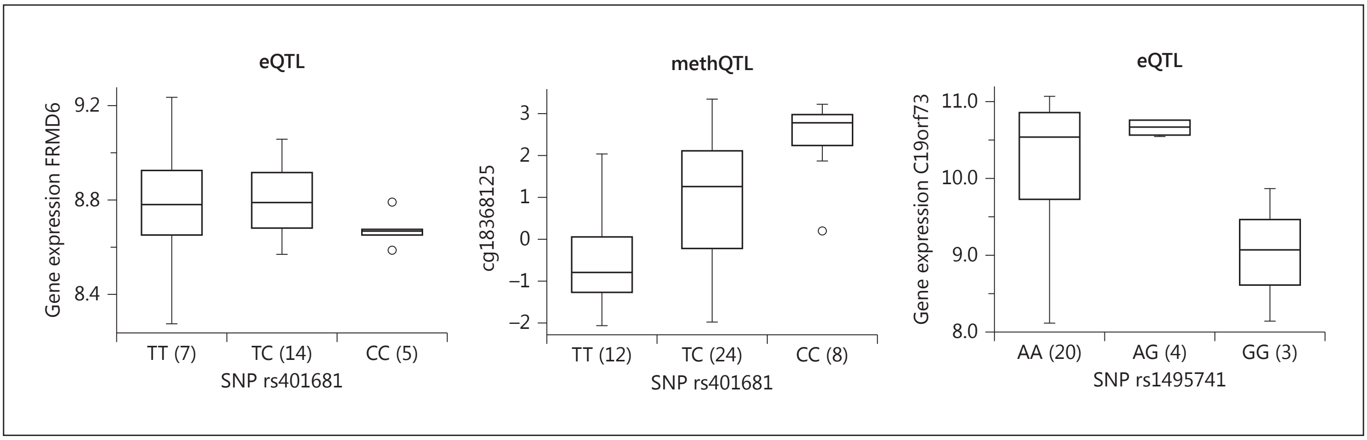


Fig. 2. GWAS-reported SNPs significantly associated with gene expression levels and/or DNA methylation levels in UBC. Values in parentheses are number of individuals with each genotype.

Table 7. Distribution of the 1,469 triple relationship directions per pair-wise analysis

eQTL	methQTL	Expression-methylation	All triplets, n (%)[a]	Pairs in *cis*-effect, n (%)[b]
+	+	+	419 (29)	1 (5)
–	–	–	58 (4)	3 (16)
+	–	–	276 (19)	4 (21)
–	+	+	78 (5)	1 (5)
–	+	–	262 (18)	6 (32)
+	–	+	62 (4)	3 (16)
–	–	+	250 (17)	1 (5)
+	+	–	64 (4)	0 (0)

[a] The total distribution for the 1,469 triplets. [b] The distribution only for the ones that had one pair in *cis*-effect.

When we checked how the significant findings are distributed in terms of the direction of the relationship, there were more QTLs positively than negatively associated (60 vs. 40% eQTL, 63 vs. 37% methQTLs), implying that having more copies of the rare allele increases the levels of the gene expression or the levels of methylation.

Lastly, for the QTL associations in our study, we investigated how many of the SNPs involved have been previously reported as a trait-associated SNPs for UBC. We found that SNP rs401681-*TERT/CLPTM1L* on chromosome 5 has been associated with the expression of *FRMD6* located on chromosome 14 (p value = 3.7×10^{-5}), and with cg18368125-*TMED6* on chromosome 16 (p value = 4.8×10^{-5}). Also, SNP rs1495741-*NAT2* on chromosome 8 has been associated with the expression of *C19orf73* located on chromosome 19 (fig. 2).

Integration of Results Derived from the Pair-Wise Analysis

From the final subset of eQTLs and methQTLs, we obtained 49,708 common SNPs (50% from the total SNPs for eQTLs and methQTLs), affecting a total of 227,572 eQTLs (207 *cis*-acting) and 298,869 methQTLs (247 *cis*-acting). Multiple expression probes and CpGs were significantly associated with more than one SNP and vice versa. We found that 1,469 QTLs belonged to a triple relationship (SNP-CpG-gene expression; see online suppl. table 4). Regarding the association patterns, the majority (29%) of these 1,469 triplets showed a positive association pattern, i.e. the higher the methylation, the higher the expression, where the rare allele is classified with higher expression and methylation levels. A second pattern (19%) was 'the higher the methylation, the lower the expression', where the rare allele is associated with high expression levels and low methylation levels. When restricted to *cis*-relationship, no triplets were found, but there were 19 pairs (1 eQTL, 1 methQTL and 17 CpG-gene expression pairs) that were in *cis*. The distribution of these pairs was completely different from that of the rest of the triplets. The most frequent pattern (32%) observed was a positive association between SNPs and methylation and a negative association of both SNPs and CpGs with the expression. All of the possible patterns with their percentages are shown in table 7. Lastly, we checked for the hotspots in these triplets and found some of them for SNPs, CpGs and gene expression probes (fig. 3).

Color version available online

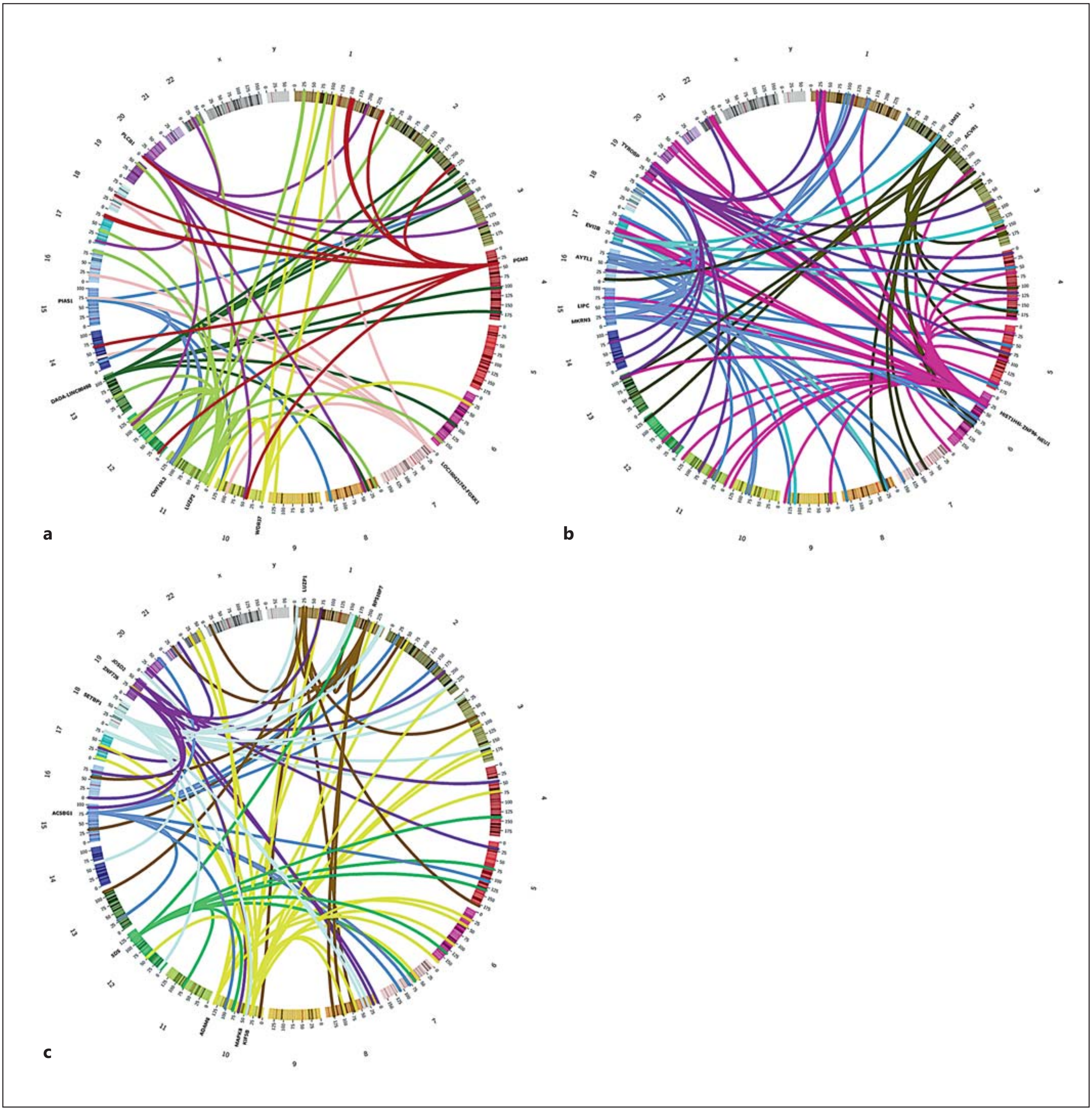


Fig. 3. Circular representation of the hotspots found for SNPs (**a**), CpGs (**b**) and gene expression probes (**c**) extracted from the relationships in the triplets. Each chromosome is represented with a different colour, and the colours of the lines correspond to the SNPs, CpGs or gene expression probes that are located on the chromosome they share the colour with. The names of the genes are located on the gene with the hotspot.

Discussion

The post-genome era delivers a wealth of omics data allowing to explore the relationships between genetics, epigenetics and gene expression, being of great importance to better understand the biological mechanism(s) underlying a disease. In the field of cancer, this integrative approach becomes particularly crucial on the basis of the knowledge indicating that SNPs, CpGs and gene expression play an important role in the development of these complex diseases [27, 28].

In this work, we propose an omics integrative analytical framework based on a multi-staged strategy, and we apply it to explore the relationships between three sets of data measured at a genome-wide level in UBC tumour samples. We provide further evidence on how common genetic variation and DNA methylation are statistically associated with the regulation of gene expression. Based on the knowledge that DNA is looped, allowing the interaction between two DNA regions located far away from each other, we did not only study *cis*- but also *trans*-relationships [29]. Here, we showed that some SNPs are associated with DNA methylation, that the latter is associated with gene expression and that some SNPs associate with both DNA methylation and gene expression.

Individual and Pair-Wise Analyses

The global pattern for methylation observed in our study (online suppl. fig. 1) parallels that reported previously for germline (blood) [14]. Consistently with previous studies performed on blood [14, 16] and human brain samples [13], we found that – when located in an island/shore – the correlations between DNA methylation and gene expression from the same gene are predominantly negative, supporting the known biological mechanisms of gene regulation (80%). DNA methylation occurs near the transcription start site of a gene, blocking the initiation of gene expression (for a review, see Jones [3]). To highlight the relevant results, four different CpGs (cg01354473, cg07778029, cg25047280, and cg26521404) located in a CpG island of the *HOXA9* gene on chromosome 8 were negatively correlated with the expression of the gene. It was reported that *HOXA9* acts as a tumour suppressor gene in oral cancer [30], while methylation of this gene has been associated with the regulation of its expression in UBC [31] and with a risk of different cancers such as breast [32], oral cavity [33] and ovarian cancer [34] as well as with a risk of recurrence in UBC [35]. The observed negative association between four CpGs and *HOXA9* expression in our study suggests that the inhibition of *HOXA9* expression may affect the development of UBC and supports the approach applied in this study.

On the other hand, the ENCODE Project provided some clues to the understanding of the biological behaviour of *trans*-relationships and of CpGs belonging to *cis*-relationships when located in a different gene [36]. In our study, we mainly observed positive correlations (79%) in all of these scenarios, meaning that increasing levels of methylation correlated with increasing levels of gene expression or the other way around, suggesting either a direct mechanism or an indirect mechanism, where methylation affects expression of a gene repressor, thus leading to an apparent association with increased gene levels. These results warrant further mechanistic studies explaining the complex association between DNA methylation and gene expression.

Little is known about the relationship between genetic variants and DNA methylation. Heyn et al. [1] has recently published a methQTL analysis using the cancer genome atlas data but only with SNPs detected in GWAS and *cis*-acting methQTLs. They detected one methQTL in UBC, where SNP rs401681 in *TERT_CLPTM1L* was associated with cg06550200 located in *CLPTM1L*; unfortunately, we have not been able to replicate this association as this CpG is not present in the 27K methylation array. Nonetheless, for the first time, we have performed *cis*- and *trans*-acting methQTL analyses in UBC tumour tissue samples using CpGs that have previously been correlated with gene expression. From this assessment, we found 538 *cis*-relationships (listed in the online suppl. table 3 with all necessary information for further studies and validation). More frequently, *cis*-relationships between genetic variants and gene expression levels have been assessed. We also performed eQTL association studies in *cis*- and *trans*-relationships under the same conditions as for methQTLs and found 441 *cis*-eQTLs (online suppl. table 2). We performed these analyses on significant expression-methylation-correlated probes identified in the first step upon the assumption that epigenetics interferes with the gene expression levels.

The proportion of eQTLs (0.01%, 471,818) and methQTLs (0.01%, 643,477) was similar, although more SNPs were involved in eQTLs (32.6%, 154,203) than in methQTLs (22.7%, 148,528), possibly because of the smaller sample size of the former. Similarly, we found no major differences in the percentages of QTL associations classified as *cis*-, *trans*- and *trans-out* according to the genomic distances defined before. Nevertheless, when considering the MAF distribution, a higher number of QTLs were observed for SNPs with a MAF of ≤0.2. While these results should be interpreted cautiously, due to the possibility of false positives, it is worth highlighting that we found a greater number of positive than negative QTL relationships, meaning that having the rare allele is associated with increased gene expression or methylation levels.

Some studies have related SNPs associated with complex diseases at genome-wide significance level to gene expression or methylation levels [1, 10, 37]. Out of the 14 GWAS UBC SNPs [38], 2 were shown to be associated with gene expression and methylation in *trans*-relationships (fig. 2). Interestingly, rs401681-*TERT/CPTL1M*, a variant strongly associated with low-grade and low-risk UBC [38], was found to be associated with a lower expression of *FRMD6* in our study, a gene that was reported to be involved in the inhibition of proliferation in human cells [39].

Integrative Analysis

We observed an enrichment of significant associations of genetic variants with methylation and gene expression with 49,708 SNPs related to 227,572 eQTLs and 298,869 methQTLs (207 eQTLs and 247 methQTLs in *cis*-relationship) suggesting a co-regulated expression and methylation. The percentage of enrichment associated with eQTLs (11.5%) and methQTLs (10.0%) was similar to that found by Wagner et al. [40] who detected an enrichment of 9.5% in fibroblasts. Bell et al. [14] also found an enrichment in lymphoblastoid cell lines. In contrast, Gibbs et al. [12] found only a modest overlap between both data in brain tissues, while Drong et al. [15] found no enrichment in adipose tissue. This highlights the fact that specific genetic variants may show tissue-specific effects and that little is known about them at a genome-wide level.

We also found a total of 1,469 QTLs, where the same SNP was significantly associated with both eQTL and methQTL in previously identified significant gene expression-CpG pairs. This three-way type relationship between SNP-CpG-gene expression supports the notion that the three data sets implemented in this study are closely related in regulating part of the genome, an observation that may provide new insight into the genetics of this complex disease. Furthermore, we observed that the most frequent pattern (29%) in these three-way relationships is a positive association pattern, suggesting that hypermethylation may act through a direct mechanisms or affect a repressor gene associated with an over-expression of gene levels. In addition, having the rare allele is associated with hypermethylation and over-expression pattern. This finding together with the fact that, in our study, 82% of the CpGs that are related with gene expression in *trans*-effect are positively correlated suggest that if one SNP is co-regulating both, this relation should be positive. Thus, we could hypothesize that the rare allele of the SNP associates with hypermethylation and, at the same time, associates with over-expression, as a possible regulation scenario in *trans*-effect.

When inspecting the *cis*-relationships, no triplets were found, but there were 19 pairs (1 eQTL, 1 methQTL and 17 CpG-gene expression pairs). In this scenario, the most frequent pattern (32%) suggests that having the rare allele is associated with hypermethylation and under-expression, where expression and methylation are associated inversely. This fact suggests another possible regulation scenario based on previous findings. We demonstrated that 79% of the CpGs located in the promoter region of the gene are negatively correlated in *cis*-relationships with the gene expression levels; meaning that higher methylation levels may effect a decrease in the gene expression levels. An example of this scenario is shown in figure 4, where SNP rs289516 located in gene *DLC1* is negatively associated in *trans*-relationship with the expression of *HOXA9* ($\beta = -1.1$; p value = 3.7×10^{-5}) and positively with cg01354473 located in the island of the *HOXA9* gene ($\beta = 1.8$; p value = 9.9×10^{-5}). The relationship between the expression and the methylation levels in the *HOXA9* gene has already been reported as negatively correlated ($r^2 = -0.7$; p value = 1.4×10^{-5}). It has been already published that the methylation of *HOXA9* is negatively correlated with the gene expression in UBC [31], as we observed in our study. We added a new step to this complex scenario, since SNP rs289516 is also involved in this triple relationship. This SNP belongs to the *DLC1* gene considered as a tumoor suppressor gene, and this particular SNP has been picked up in two GWAS, one for asthma [41] and one for breast cancer [42], but any of them passed the GWAS significant threshold. Other examples with biological support are the triplet composed by SNP rs29658399 located on gene *DNAH11*, the gene expression of *HSPA1A* and cg00929855 located on gene *HSPA1A*. It has previously been published that the *HSPA1A* promoter methylation underlies the defect in gene expression reduction observed in UBC cell lines [43]. In addition we found some hotspots in these triplets regarding SNPs, CpGs and gene expressions probes.

In the Circos plot (fig. 3a), we observed a predominant relation for one SNP (rs10569 located on gene *PGM2*) on chromosome 4. *PGM2* is a protein-coding gene and is as-

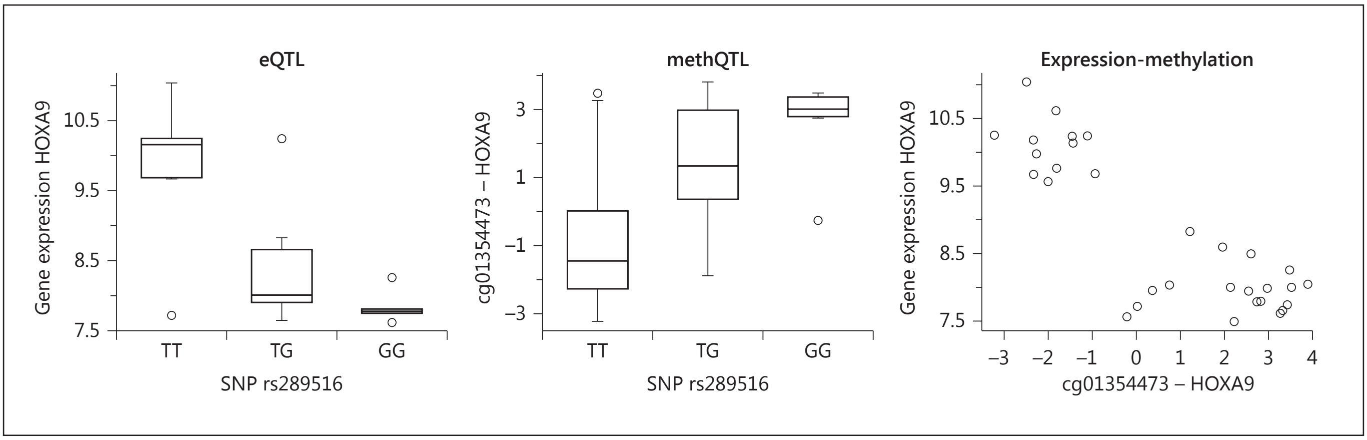


Fig. 4. Example of one triple relationship where integrated common genetic variants with DNA methylation and gene expression is one of the main possible scenarios for regulation.

sociated with diseases such as pneumonia and hypoxia. While alterations in this gene have not yet been directly associated with cancer, hypoxia is a known relevant process for tumour survival. This SNP has been positively associated with the expression of *SETBP1,* coding for an important cancer gene located on chromosome 18 that is observed also as a predominant hotspot in figure 3c. Somatic mutations in *SETBP1* [44], as well as its expression patterns [45], are related with myeloid leukaemia disease. Moreover in figure 3b, we observed a very predominant hotspot regarding three CpGs belonging to three different genes but all closely located on chromosome 6; Two of them (cg02622316 located on gene *ZNF96* and cg02599464 located on gene *HIST1H41*) have already been published as hypermethylated in individuals with muscle-invasive bladder cancer [46]. The first one is positively associated with many SNPs and gene expression probes, and the second is positively and negatively associated with some SNPs and only positively with some gene expression probes.

A more detailed discussion of the potential biological findings than involved in triple relationships is beyond this particularly study and detailed results about all of the combinations are provided in online supplementary table 4.

The Integrative Framework

We built and propose a multi-staged omics integration framework whose application does not require a strong methodological knowledge, being easy and effective to use. The multi-staged framework we applied has the advantage of analyzing data of all subjects that overlap among pairs of data and is not restricted to only those few individuals with a complete overlap among all data types. Thus, we take advantage of more samples using this framework than integrating the data in a multi-dimensional model.

Therefore, we show here, for the first time, the application of a multi-staged framework that allows us to (1) integrate more than two omics data for the same set of individuals; (2) dissect the biological relationships that may point to new mechanisms involved in the development/progression of UBC through a hypothesis-based model built step by step, and (3) envision the complexities of the general scenario of genomic regulation.

Conclusions

While these results are exciting, we acknowledge the following limitations. First, in this study, we use the 27K methylation array that only covers a selection of CpG sites making it infeasible to replicate previous reported findings using the 450K array. Second, statistical power is a commonplace in any QTL analysis, given the extensive amount of data analyzed and the small sample size. While this limitation needs to be considered in the interpretation of the results, it is worth mentioning that even a large enough size will unlikely be able to meet the standard criteria of statistical power; therefore, our study represents a proof of concept in the integrative omics field. In addition, while we might not be able to address unmeasured confounding factors, no differences were found between the demographic factors and methylation and gene expression in our series. A validation of these results, to discard false positive findings, is not trivial due to the multiple genomic factors, the models considered and the characteristics of the series.

Despite these limitations, this study has several strengths. We have performed an analysis of tumour samples what gave us the opportunity to study in detail the regulation of three types of omics data in UBC, providing some evidences on the genomics regulation of the tumour. We have applied an easy, reproducible and detailed framework to perform an integrative study of the relationships between genetic variations, DNA methylation and gene expression, showing a whole spectrum of associations between them. We have shown that omics data integration helps unravelling biological mechanisms involved in UBC. All of these relations may help in the identification of new molecular targets to be further explored in detail, mainly regarding *trans*-relationships.

In conclusion, this study provides the scientific community with a pipeline to integrate more than two sets of omics data that can be applied in future analyses seeking to better understand the biology behind complex diseases. In addition, we highlight the importance of integrating omics data to identify new genetic mechanisms in UBC. While several pieces of evidences support our findings, they still require experimental validation to be considered conclusive.

Acknowledgements

We are grateful to Stephen J. Chanock (Division of Cancer Epidemiology and Genetics, National Cancer Institute, Department of Health and Human Services, Bethesda, Md., USA) and Alfredo Carrato (Hospital Ramón y Cajal, Madrid, Spain) for scientific input to the work and to the coordinators, field and administrative workers, technicians, and patients of the pilot Spanish Bladder Cancer (SBC)/EPICURO Study. This work has partially been funded by Obra Social Fundación 'la Caixa'; Red Temática de Investigación Cooperativa en Cáncer (RTICC, No. RD12/0036/0050) and Fondo de Investigaciones Sanitarias (FIS, No. PI06-1614), Instituto de Salud Carlos III (ISCIII), the Spanish Ministry of Economy and Competitiveness; the Intramural Research Program of the National Institutes of Health, National Cancer Institute, Division of Cancer Epidemiology and Genetics (contract NCI NO2-CP-11015); the Belgian Science Policy Office Phase VII IAP network 'Dynamical systems, control and optimization' (DYSCO II). The Fonds de la Recherche Scientifique (FNRS), in particular the F.R.F.C. PDR convention No. 2.4609.11, in part supported the contributions of K.B. and K.V.S.

References

1 Heyn H, Sayols S, Moutinho C, et al: Linkage of DNA methylation quantitative trait loci to human cancer risk. Cell Rep 2014;7:331–338.
2 Irizarry RA, Ladd-Acosta C, Wen B, et al: The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet 2009;41:178–186.
3 Jones PA: Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 2012;13:484–492.
4 Portela A, Esteller M: Epigenetic modifications and human disease. Nat Biotechnol 2010;28:1057–1068.
5 Esteller M: Epigenetics in cancer. N Engl J Med 2008;358:1148–1159.
6 Cheung VG, Spielman RS: Genetics of human gene expression: mapping DNA variants that influence gene expression. Nat Rev Genet 2009;10:595–604.
7 Nica AC, Montgomery SB, Dimas AS, et al: Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet 2010;6:e1000895.
8 Nicolae DL, Gamazon E, Zhang W, et al: Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet 2010;6:e1000888.
9 Pickrell JK, Marioni JC, Pai AA, et al: Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 2010;464:768–772.
10 Westra HJ, Peters MJ, Esko T, et al: Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet 2013;45:1238–1243.
11 Zhernakova DV, de Klerk E, Westra H-J, et al: DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. PLoS Genet 2013;9:e1003594.
12 Gibbs JR, van der Brug MP, Hernandez DG, et al: Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet 2010;6:e1000952.
13 Zhang D, Cheng L, Badner JA, et al: Genetic control of individual differences in gene-specific methylation in human brain. Am J Hum Genet 2010;86:411–419.
14 Bell JT, Pai AA, Pickrell JK, et al: DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol 2011;12:R10.
15 Drong AW, Nicholson G, Hedman AK, et al: The presence of methylation quantitative trait loci indicates a direct genetic influence on the level of DNA methylation in adipose tissue. PLoS One 2013;8:e55923.
16 Van Eijk KR, de Jong S, Boks MPM,et al: Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. BMC Genomics 2012;13:636.
17 Liu Y, Ding J, Reynolds LM, et al: Methylomics of gene expression in human monocytes. Hum Mol Genet 2013;22:5065–5074.
18 Li Q, Seo JH, Stranger B, et al: Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. Cell 2013;152:633–641.
19 Greenawalt DM, Sieberts SK, Cornelis MC, et al: Integrating genetic association, genetics of gene expression, and single nucleotide polymorphism set analysis to identify susceptibility loci for type 2 diabetes mellitus. Am J Epidemiol 2012;176:423–430.
20 Bibikova M, Le J, Barnes B, et al: Genome-wide DNA methylation profiling using Infinium(R) assay. Epigenomics 2009;1:177–200.
21 Du P, Zhang X, Huang C-C, et al: Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 2010;11:587.
22 Chen Y, Choufani S, Ferreira JC, et al: Sequence overlap between autosomal and sex-linked probes on the Illumina HumanMethylation27 microarray. Genomics 2011;97:214–222.

DOI: 10.1159/000381184

23 Kauffmann A, Gentleman R, Huber W: arrayQualityMetrics – a bioconductor package for quality assessment of microarray data. Bioinformatics 2009;25:415–416.

24 Gautier L, Cope L, Bolstad BM, Irizarry RA: affy – analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 2004;20:307–315.

25 Benjamini Y, Yekutieli D: The control of the false discovery rate in multiple testing under dependency. Ann Stat 2001;29:1165–1188.

26 Krzywinski M, Schein J, Birol I, et al: Circos: an information aesthetic for comparative genomics. Genome Res 2009;19:1639–1645.

27 You JS, Jones PA: Cancer genetics and epigenetics: two sides of the same coin? Cancer Cell 2012;22:9–20.

28 Kanwal R, Gupta S: Epigenetic modifications in cancer. Clin Genet 2012;81:303–311.

29 Bickmore WA, van Steensel B: Genome architecture: domain organization of interphase chromosomes. Cell 2013;152:1270–1284.

30 Uchida K, Veeramachaneni R, Huey B, et al: Investigation of HOXA9 promoter methylation as a biomarker to distinguish oral cancer patients at low risk of neck metastasis. BMC Cancer 2014;14:353.

31 Reinert T, Modin C, Castano FM, et al: Comprehensive genome methylation analysis in bladder cancer: identification and validation of novel methylated genes and application of these as urinary tumor markers. Clin Cancer Res 2011;17:5582–5592.

32 Gilbert PM, Mouw JK, Unger MA, et al: HOXA9 regulates BRCA1 expression to modulate human breast tumor phenotype. J Clin Invest 2010;120:1535–1550.

33 Guerrero-Preston R, Soudry E, Acero J, et al: NID2 and HOXA9 promoter hypermethylation as biomarkers for prevention and early detection in oral cavity squamous cell carcinoma tissues and saliva. Cancer Prev Res (Phila) 2011;4:1061–1072.

34 Wu Q, Lothe RA, Ahlquist T, et al: DNA methylation profiling of ovarian carcinomas and their in vitro models identifies HOXA9, HOXB5, SCGB3A1, and CRABP1 as novel targets. Mol Cancer 2007;6:45.

35 Reinert T, Borre M, Christiansen A, et al: Diagnosis of bladder cancer recurrence based on urinary levels of EOMES, HOXA9, POU4F2, TWIST1, VIM, and ZNF154 hypermethylation. PLoS One 2012;7:e46297.

36 ENCODE Project Consortium: The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 2004;306:636–640.

37 Fu Y-P, Kohaar I, Rothman N, et al: Common genetic variants in the PSCA gene influence gene expression and bladder cancer risk. Proc Natl Acad Sci USA 2012;109:4974–4979.

38 Rothman N, Garcia-Closas M, Chatterjee N, et al: A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. Nat Genet 2010;42:978–984.

39 Visser-Grieve S, Hao Y, Yang X: Human homolog of Drosophila expanded, hEx, functions as a putative tumor suppressor in human cancer cell lines independently of the Hippo pathway. Oncogene 2012;31:1189–1195.

40 Wagner JR, Busche S, Ge B, et al: The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome Biol 2014;15:R37.

41 Moffatt MF, Gut IG, Demenais F, et al: A large-scale, consortium-based genomewide association study of asthma. N Engl J Med 2010;363:1211–1221.

42 Hunter DJ, Kraft P, Jacobs KB, et al: A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 2007;39:870–874.

43 Qi W, White MC, Choi W, et al: Inhibition of inducible heat shock protein-70 (hsp72) enhances bortezomib-induced cell death in human bladder cancer cells. PLoS One 2013;8:e69509.

44 Piazza R, Valletta S, Winkelmann N, et al: Recurrent SETBP1 mutations in atypical chronic myeloid leukemia. Nat Genet 2013;45:18–24.

45 Makishima H, Yoshida K, Nguyen N, et al: Somatic SETBP1 mutations in myeloid malignancies. Nat Genet 2013;45:942–946.

46 Ibragimova I, Dulaimi E, Slifker MJ, et al: A global profile of gene promoter methylation in treatment-naïve urothelial cancer. Epigenetics 2014;9:760–773.

PLOS | ONE

Genetic Variation in the *TP53* Pathway and Bladder Cancer Risk. A Comprehensive Analysis

Silvia Pineda[1], Roger L. Milne[1], M. Luz Calle[2], Nathaniel Rothman[3], Evangelina López de Maturana[1], Jesús Herranz[1], Manolis Kogevinas[4,5], Stephen J. Chanock[3], Adonina Tardón[6], Mirari Márquez[1], Lin T. Guey[1], Montserrat García-Closas[3], Josep Lloreta[5,7], Erin Baum[1], Anna González-Neira[1], Alfredo Carrato[8,9], Arcadi Navarro[10,11,12,13], Debra T. Silverman[3], Francisco X. Real[1,10], Núria Malats[1]*

1 Spanish National Cancer Research Center (CNIO), Madrid, Spain, **2** Systems Biology Department, University of Vic, Vic, Spain, **3** Division of Cancer Epidemiology and Genetics, National Cancer Institute, Department of Health and Human Services, Bethesda, Maryland, United States of America, **4** Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain, **5** Institut Municipal d'Investigació Mèdica – Hospital del Mar, Barcelona, Spain, **6** Department of Preventive Medicine, Universidad de Oviedo, Oviedo, Spain, **7** Departament de Patologia, Hospital del Mar – IMAS, Barcelona, Spain, **8** Servicio de Oncología, Hospital Universitario de Elche, Elche, Spain, **9** Servicio de Oncología, Hospital Universitario Ramon y Cajal, Madrid, Spain, **10** Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain, **11** Institut de Biologia Evolutiva (UPF-CSIC), Barcelona, Spain, **12** Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain, **13** Instituto Nacional de Bioinformática, Barcelona, Spain

Abstract

Introduction: Germline variants in *TP63* have been consistently associated with several tumors, including bladder cancer, indicating the importance of *TP53* pathway in cancer genetic susceptibility. However, variants in other related genes, including *TP53* rs1042522 (Arg72Pro), still present controversial results. We carried out an in depth assessment of associations between common germline variants in the *TP53* pathway and bladder cancer risk.

Material and Methods: We investigated 184 tagSNPs from 18 genes in 1,058 cases and 1,138 controls from the Spanish Bladder Cancer/EPICURO Study. Cases were newly-diagnosed bladder cancer patients during 1998–2001. Hospital controls were age-gender, and area matched to cases. SNPs were genotyped in blood DNA using Illumina Golden Gate and TaqMan assays. Cases were subphenotyped according to stage/grade and tumor p53 expression. We applied classical tests to assess individual SNP associations and the Least Absolute Shrinkage and Selection Operator (LASSO)-penalized logistic regression analysis to assess multiple SNPs simultaneously.

Results: Based on classical analyses, SNPs in *BAK1* (1), *IGF1R* (5), *P53AIP1* (1), *PMAIP1* (2), *SERINPB5* (3), *TP63* (3), and *TP73* (1) showed significant associations at p-value≤0.05. However, no evidence of association, either with overall risk or with specific disease subtypes, was observed after correction for multiple testing (p-value≥0.8). LASSO selected the SNP rs6567355 in *SERPINB5* with 83% of reproducibility. This SNP provided an OR = 1.21, 95%CI 1.05–1.38, p-value = 0.006, and a corrected p-value = 0.5 when controlling for over-estimation.

Discussion: We found no strong evidence that common variants in the *TP53* pathway are associated with bladder cancer susceptibility. Our study suggests that it is unlikely that *TP53* Arg72Pro is implicated in the UCB in white Europeans. *SERPINB5* and *TP63* variation deserve further exploration in extended studies.

Citation: Pineda S, Milne RL, Calle ML, Rothman N, López de Maturana E, et al. (2014) Genetic Variation in the *TP53* Pathway and Bladder Cancer Risk. A Comprehensive Analysis. PLoS ONE 9(5): e89952. doi:10.1371/journal.pone.0089952

Editor: Masaru Katoh, National Cancer Center, Japan

Received October 29, 2013; **Accepted** January 24, 2014; **Published** May 12, 2014

Funding: This work was supported by the Fondo de Investigación Sanitaria, Spain (grant numbers 00/0745, PI051436, PI061614, G03/174); Red Temática de Investigación Cooperativa en Cáncer (grant number RD06/0020-RTICC), Spain; Marató TV3 (grant number 050830); European Commission (grant numbers EU-FP7-HEALTH-F2-2008-201663-UROMOL; US National Institutes of Health (grant number USA-NIH-RO1-CA089715); and the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute at the National Institutes of Health, USA; Consolíder ONCOBIO (Ministerio de Economía y Competitividad, Madrid, Spain). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: FX Real is a PLOS ONE Editorial Board member. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: nmalats@cnio.es

Introduction

In more developed countries, urothelial carcinoma of the bladder (UCB) is the fourth most common cancer in men and the seventeenth in women, the overall male:female ratio being 3:1. This ratio is greater (6:1) in Spain, where the disease presents one of the highest incidence rates among men (51 per 100,000 man-year) [1]. Tobacco smoking and occupational exposure to aromatic amines have been established as the strongest risk factors, among others [2]. While no high-penetrance allele/gene has been identified to date as associated with UCB, there is well-established evidence that UCB risk is influenced by common genetic variants [3,4].

Previous studies characterizing UCB are consistent with the existence of, at least, two disease subtypes based on their morphological and genetic features. The first subtype includes low-risk, papillary, non-muscle invasive tumors (NMIT, 60–65% of all UCB) and the second type includes both high-risk NMIT (15–20% of all UCB) and muscle invasive tumors (MIT, 20%–30% of all UCB). Supporting these morphological subtypes, differential genetic pathways were described and were associated with distinct UCB evolution. Somatic mutations in *FGFR3* are more frequent in low-risk NMIT, while mutations in *TP53* and *RB* are mainly involved in high-risk NMIT and MIT [5,6]; mutations in *PIK3CA* and *HRAS* occur similarly in the two tumor subtypes. Interestingly, an exploratory analysis has shown that some germline genetic variants might be differentially associated with the risk of developing distinct UCB subphenotypes defined according to tumor stage (T) and grade (G) [7].

TP53 is the most important human tumor suppressor gene and its implications in UCB have been extensively studied [8]. *TP53* is located in17p13, a region that is frequently deleted in human cancers, and it encodes the p53 protein. p53 is a transcription factor controlling cell proliferation, cell cycle, cell survival, and genomic integrity and - therefore - it regulates a large number of genes. Under normal cellular conditions, p53 is rapidly degraded due to the activity of *MDM2*, a negative p53 regulator that is also a p53 target gene. Upon DNA damage or other stresses, p53 is stabilized and regulates the expression of many genes involved in cell cycle arrest, apoptosis, and DNA repair among others. Somatic alterations in *TP53*/p53 are one of the most frequent alterations associated with UCB, especially with the more aggressive tumors [9].

Germline *TP53* mutations predispose to a wide spectrum of early-onset cancers and cause Li-Fraumeni and related syndromes [10,11]. These mutations are usually single-base substitutions. Over 200 germline single nucleotide polymorphisms (SNPs) in *TP53* have been identified at present [12]. SNP rs1042522 (Arg72Pro) has been assessed in association with several cancers, among them UCB. However, the results of these studies are inconsistent [13,14,15,16,17,18]. In contrast, an association between SNP rs710521 in *TP63*, a *TP53* family member, and risk of UCB has been convincingly replicated, pointing to the involvement of *TP53* pathway members in UCB susceptibility [4].

The aim of this study was to comprehensively investigate whether germline SNPs in genes involved in the *TP53* pathway are associated with risk of UCB. To this end, a total of 184 tagSNPs in 18 key genes were assessed using data from the Spanish Bladder Cancer/EPICURO study.

Materials and Methods

Study Subjects

The Spanish Bladder Cancer/EPICURO Study is a case-control study carried out in 18 hospitals from five areas in Spain and described elsewhere [2,4,7]. Briefly, cases were patients diagnosed with primary UCB at age 21–80 years between 1998 and 2001. All participants were of self-reported white European ancestry. Diagnostic slides from each patient were reviewed by a panel of expert pathologists to confirm the diagnosis and to ensure that uniform classification criteria were applied based on the 1999 World Health Organization and International Society of Urological Pathology systems [19].

Controls were patients admitted to participating hospitals for conditions thought to be unrelated to the UCB risk factors. The main reasons for hospital admission were: hernia (37%), other abdominal surgery (11%), fracture (23%), other orthopaedic problem (7%), hydrocoele (12%), circulatory disorder (4%), dermatological disorder (2%), ophthalmological disorder (1%), and other diseases (3%). Controls were individually matched to the cases on age within 5-year categories, gender, ethnic origin and region of residence.

Information on sociodemographics, smoking habits, occupational and environmental exposures, and past medical and familial history of cancer was collected by trained study monitors who conducted a comprehensive computer- assisted personal interview with the study participants during their hospital stay. Of 1,457 eligible cases and 1,465 controls, 1,219 (84%) and 1,271 (87%), were interviewed, respectively.

All subjects gave written informed consent to participate in the study, which was approved by the ethics committees of the participating centers.

Genotyping

A total of 184 tagSNPs from 18 genes participating in the *TP53* pathway were selected using the Select Your SNPs (SYSNPs) program [20]. SYSNP used information from dbSNP b25, hg17 and HapMap Release #21. Haploview's Tagger algorithm (v3.32) was applied with default parameter values. The tool considers all available information for each SNP and implements algorithms that provide the status of each SNP as a tagSNP, a captured SNP or a non-captured SNP. According to this information tagSNPs were selected. The following groups of genes were considered: 1) *TP53* family members (*TP53*, *TP63* and *TP73*) and 2) genes known to be targets of p53 or regulators of p53 function [*BAK1, BAX, BBC3, BIRC5, CDKN1A, FAS, GADD45A, IGF1R, MDM2, PCNA, PMAIP1, SERPINB5, SFN* (Stratifin, 14-3-3sigma), *TP53AIP1*), and 3) c-*MYC*, a major oncogene involved in a broad range of human cancers that regulates p53 pro-apoptotic activity (See Table S1 in File S1). SNPs were genotyped using Illumina Golden Gate and TaqMan (Applied Biosystems) assays at the Spanish Core Genotyping Facility at the CNIO (CEGEN- CNIO). Genotyping was successful for 1,058 cases and 1,138 controls. We calculated the coverage for each gene using Haploview 4.2 by selecting the SNPs within a gene with a MAF≥0.05 from the 1000 genomes project, as reference, and obtained the number of SNPs captured with the SNPs genotyped at r2≥0.8 within each gene.

Statistical Analysis

Departure from Hardy-Weinberg equilibrium was assessed in controls using Pearson's chi-squared test. Missing genotypes were imputed for the multi-SNP model using the BEAGLE 3.0 method [21]. Associations between UCB and the SNPs considered were assessed using two approaches: classical logistic and polytomous regression analyses applied to each SNP individually, and the Least Absolute Shrinkage and Selection Operator (LASSO)-penalized logistic regression to assess all SNPs simultaneously. All models were adjusted for age at diagnosis (cases) or interview (controls), gender, region, and smoking status. Smoking status was coded in four categories (never: <100 cigarettes in their lifetime; occasional: at least one per day for ≥6 months; former: if they had smoked regularly, but stopped at least 1 year before the study inclusion date; and current: if they had smoked regularly within a year of the inclusion date [2].

With the "classical" statistical approaches we assessed SNP main effects for the whole disease and for different subtypes of UCB, as well as SNP*SNP and SNP*smoking interactions. Disease subtypes were defined in two ways. First, according to established criteria based on tumor stage (T) and grade (G) as low-risk NMIT (TaG1 and TaG2), high-risk NMIT (TaG3, T1G2, T1G3, and Tis), and MIT (T2, T3, and T4); and second, according to the

Table 1. Demographics and smoking status of patients included in the study.

	Cases (n = 1058)	Controls (n = 1138)	[1]*p-value*
Gender			
Male	920 (87%)	991 (87%)	
Female	138 (13%)	147 (13%)	0.9
Age			
<55	149 (14%)	181 (16%)	
55–64	222 (21%)	278 (24%)	
65–69	241 (23%)	263 (23%)	
70–74	225 (21%)	222 (20%)	
75+	221 (21%)	194 (17%)	0.06
Region			
1-Barcelona	214 (20%)	233 (21%)	
2-Valles	173 (16%)	181 (16%)	
3-Elche	83 (8%)	80 (7%)	
4-Tenerife	195 (19%)	207 (18%)	
5-Asturias	393 (37%)	437 (38%)	0.9
Smoking			
Never	147 (14%)	334 (29%)	
Occasional	43 (4%)	81 (7%)	
Former	409 (38%)	429 (38%)	
Current	454 (43%)	283 (25%)	<0.001
Missing	5 (1%)	11 (1%)	

[1]*p-value* from Pearson's χ^2 test for association.

doi:10.1371/journal.pone.0089952.t001

tumor expression of p53 determined using DO7 antibody. We applied the histoscore as $z = \sum_{i=1}^{3} i * pos\%cells_i$, where $pos\%cells_i$ was the percentage of cells with intensity $i(i = 1,2,3)$. We then classified cases as having low or high p53 expression relative to the median histoscore.

To assess overall main effects, the four modes of inheritance were considered: co- dominant, dominant, recessive, and additive. The statistical significance of associations was determined using the Likelihood Ratio Test (LRT). We evaluated associations between individual SNPs and subtypes of UCB using polytomous logistic regression. Heterogeneity by disease subtype was tested by a LRT comparing this model to that with the ln(OR) restricted to be equal across subtypes. We also evaluated all two-way interactions between SNPs by a LRT comparing logistic regression models with the two SNPs (additive model) and covariates described above, with and without a single interaction term for multiplicative, per-allele effects. Interactions between each SNP and cigarette use (never vs. ever) were assessed using a similar method. Multiple testing was accounted for by applying a permutation test with 1,000 replicates. We applied Quanto (http://hydra.usc.edu/gxe/) to assess statistical power considering the available sample size.

We also assessed combined SNP effects using LASSO. The method has been described in detail by [22]. Briefly, the log-likelihood function applied in classical logistic regression

$$L_n(\beta) \sum_{i=1}^{n} [\mathrm{y}_i \log \pi(X'_i\beta) + (1-y_i)\log(1-\pi(X'_i\beta))], \qquad (1)$$

where n is the number of observations, is reconstructed incorporating a penalty so that

$$g(\beta;\lambda) = L_n(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|, \qquad (2)$$

where p is the number of SNPs and λ is the lasso penalty. The Newton-Raphson algorithm is applied to equation (2) to estimate β's in an iterative way.

The LASSO method is based on the idea of removing irrelevant predictor variables ($\beta = 0$) via the penalty parameter, thereby selecting only the most relevant SNPs as the subset of markers most associated with the disease. The application of the penalty parameter also avoids overfitting due to both high-dimensionality and collinearity between covariates. We only considered additive genetic mode of inheritance.

This technique gives biased estimators to reduce their variance. Because of this, the implemented package in R does not provide estimates p-values for the regression beta coefficients, since standard errors are not meaningful under a biased estimator. We therefore evaluated the results by first applying the LASSO using a 5-fold cross-validation (CV) method [23] to choose the optimal λ as that giving the minimum Akaike information criterion (AIC); we then selected the subset of SNPs that were most informative with that λ. We assessed the robustness of each SNP selected in the optimal model by calculating the reproducibility as the proportion of times each SNP was selected to be in the multivariate model from 1,000 bootstrap subsamples [24].

To evaluate the association with UCB risk of that subset of SNPs, we tested them by the LRT in a multivariate regression

Table 2. SNPs in *TP53* and bladder cancer risk.

	Cases			Controls			Additive model			Co-dominant model							
SNP	**AA**	**Aa**	**aa**	**AA**	**Aa**	**aa**	**OR**	**95% CI**	***p-value***	**OR(Aa)**	**95% CI**	***p-value***	**OR(aa)**	**95% CI**	***p-value***	**P-trend**	**Repr. (%)**
rs1042522[1]	588	372	72	628	388	84	1.04	0.91–1.20	0.5	1.10	0.91–1.33	0.3	0.97	0.68–1.37	0.8	0.5	24%
rs12951053	915	109	3	972	122	5	0.98	0.75–1.27	0.9	1.04	0.79–1.39	0.7	0.64	0.14–2.88	0.5	0.8	35%
rs1625895	761	241	28	793	266	26	1.04	0.88–1.24	0.6	0.99	0.80–1.21	0.9	1.28	0.73–2.26	0.4	0.7	13%
rs2287497	835	183	9	869	207	11	0.95	0.77–1.17	0.7	0.97	0.77–1.22	0.8	0.70	0.28–1.74	0.4	0.7	48%
rs2909430	749	251	28	800	272	27	1.03	0.87–1.23	0.7	1.04	0.85–1.27	0.7	1.23	0.70–2.16	0.5	0.7	36%
rs8073498	425	467	132	435	521	128	0.99	0.87–1.13	0.9	0.94	0.78–1.13	0.5	1.01	0.75–1.34	0.9	0.8	44%
rs8079544	923	103	2	993	102	4	1.05	0.79–1.39	0.7	1.10	0.81–1.48	0.5	0.42	0.07–2.33	0.3	0.5	40%

Repr. (%),percentage reproducibility assessing the robustness of each SNP by LASSO.
AA, Aa and aa represent common-homozygotes, heterozygotes and rare-allele homozygotes, respectively.
OR, odds ratio; CI, confidence interval; OR(Aa) and OR(aa) were estimated relative to genotype AA.
[1]Arg72Pro polymorphism.
All models were adjusted for age, gender, region and cigarette smoking status.
doi:10.1371/journal.pone.0089952.t002

model with all the SNPs in comparison to the null model. To correct for the over-estimation due the pre-selection of the best SNPs, we performed a permutation test with 10,000 replicates.

STATA 10 was used to run the classical logistic and multinomial regression analyses. All other statistical analyses were run in R (http://www.R-project.org), using the penalized library [25] for LASSO penalized logistic regression.

Results

Table 1 shows the distribution of the study subjects included in the analysis: 1,058 cases and 1,138 controls. Most individuals (87%) were male and cases were more likely to be current smokers than controls (43% vs. 25%, respectively, p-value<0.001).

No evidence of departure from Hardy-Weinberg equilibrium was observed for any SNPs after consideration of multiple testing (unadjusted p-value$>10^{-4}$). Polymorphisms in *TP53* were not individually associated with UCB risk, even at a nominal, uncorrected 5% significance level (uncorrected p-value>0.4). The percentage of reproducibility from the LASSO model using 1,000 bootstrap subsamples was <50%, indicating a poor robustness of the models. Results for the additive and co-dominant models are summarized in Table 2.

Using classical logistic regression, SNPs in *BAK1* (1), *IGF1R* (5), *P53AIP1* (1), *PMAIP1* (2), *SERPINB5* (3), *TP63* (3), and *TP73* (1) showed significant results, at a non-corrected p-value≤0.05, with overall UCB risk (Table 3). However, no evidence of association with risk was observed for any individual SNPs after correcting for multiple testing (permutation test p-value>0.8). This was also the case for the associations with the established disease subtypes defined according to stage/grade or by p53 expression (Figure 1). Of note, SNPs rs3758483 and rs983751 in *FAS* were differentially and inversely associated with MIT and high p53 expressing tumors in uncorrected analyses (Tables S2 and S3 in File S1). We also observed no evidence of SNP*SNP interactions or interactions between SNPs and smoking status (data not shown).

When all 184 SNPs were simultaneously assessed using LASSO, the method selected rs6567355 in *SERPINB5* with a reproducibility = 83%. This SNP provided an OR = 1.21, 95%CI 1.05–1.38, p-value = 0.006 in the main effect logistic regression model and a corrected p-value = 0.5 when controlling for over-estimation (Table 3). While not selected by LASSO in the last model under the stringent criteria applied, *IGF1R*-rs1058696 (OR = 0.63, 95%CI 0.44–0.90, p-value = 0.010) and *TP63*-rs13321831 (OR = 1.36, 95%CI 1.06–1.73, p-value = 0.014) showed a percentage of reproducibility >80%.

Discussion

We genotyped common variants in genes in the *TP53* pathway in 1,058 cases and 1,138 controls of white European ancestry and found no strong evidence of association with risk of UCB overall, or with subtypes of the disease defined by stage and grade or by p53 expression.

A key gene in the pathway is *TP53*, and the most commonly studied variant in this particular gene is Arg72Pro (rs1042522). Its implication in susceptibility to various cancers has been reported in Asian populations, but not in white Europeans. A meta- analysis of 49 cervical cancer studies contributing a total of 7,946 cases and 7,888 controls found that the Arg allele was associated with an increased risk of cervix cancer [14]. However, another meta-analysis of 39 studies (26,041 cases and 29,679 controls) found weak evidence for an association of the same variant with reduced breast cancer risk [18]. Regarding gastric cancer, a combined analysis of 6,859 cases and 9,277 controls from 28 studies found a

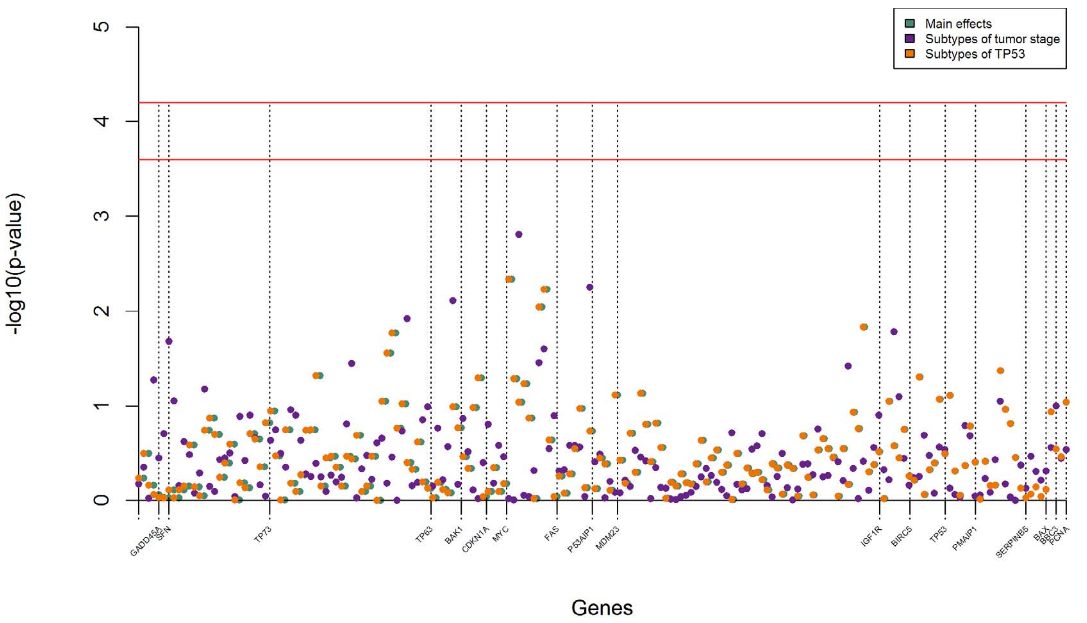


Figure 1. Main effect *p-values* for bladder cancer risk (overall and for each subphenotype) for each tag-SNP under the additive mode of inheritance. A SNP *p-value* above the red line is considered as associated with the phenotype after multiple testing correction by Bonferroni (4.2 for main effects and 3.6 for subtypes). All models are adjusted for age, gender, region and cigarette smoking status.
doi:10.1371/journal.pone.0089952.g001

stronger inverse association only among Asians [26]. For lung cancer, a marginally significant increased risk was in a combined analysis of data with 15,647 cases and 14,391 controls from 36 studies, though the association seemed to be also confined to the Asian population [27].

The association between *TP53* Arg72Pro and UCB risk has been assessed by two meta-analyses. Overall, no association was observed by Jiang et al. when comparing 1,601 cases and 1,948 controls from 10 studies, although a marginally significant association was seen among Asians (OR = 0.77, 95%CI 0.59–1.00, for ArgArg/ArgPro vs. ProPro) [13]. Discordant results have been recently reported combining data from 14 studies contributing with 2,176 cases and 2,798 controls (OR = 1.268, 95%CI 1.003–1.602, for ArgArg/ArgPro vs. ProPro among the Asian population) [17]. A large number of studies overlap between the two meta-analyses. The lack of information on gene-gene and gene-environment interactions, as well as on the concomitant effect of *TP53* somatic mutations may explain the discordant results [28].

The findings from our study confirm the lack of association of Arg72Pro in *TP53* with risk of UCB in white Europeans (OR = 0.98, 95%CI 0.77–1.26, for ArgPro vs. ArgArg and OR = 0.91, 95%CI 0.75–1.09, for ProPro vs. ArgArg, p-value = 0.5 for overall effects) [13,17]. However, we cannot rule out that lack of statistical power may hamper identification of a small effect association: even with its large sample size, the present study sample size could detect an OR≥1.3 per-allele for this SNP with 90% statistical power and at a significance level of 5%.

Regarding other SNPs in *TP53*, Lin et al reported an association with rs9895829 and rs1788227 (p-value = 0.003 and 0.027, respectively) in a smaller study with 201 cases and 311 controls in an Asian population [29]. We did not genotype these SNPs, though they are in high LD with two SNPs considered here: rs8079544 (LD = 1.0) and rs12951053 (LD = 0.7), respectively. Nonetheless, none of the assessed additional SNPs in *TP53* appeared to be associated with UCB risk. The partial coverage of the gene with the assessed SNPs (38%) does not allow us to dismiss the role of *TP53* in UCB susceptibility.

TP63 is another key member of the studied pathway. One SNP (rs710521) located in this gene has been reported to be associated with risk of UCB by a GWAS (per-allele OR = 1.19, 95%CI 1.12–1.27, p-value = 1.15×10^{-7}) [30]. This association was convincingly replicated in a combined analysis of data from different studies (allele-specific OR = 1.18, 95%CI 1.12–1.24, p-value = 1.8×10^{-10}), including ours, for which it was genotyped as part of a separate initiative [4]. Of note, this particular SNP did not show significant results in our study (OR = 0.95, 95%CI 0.83–1.10, p-value = 0.5), a fact that can be explained by the different geographical location related exposures of the participating studies, being UCB an environmental driven disease [31]. The present study assessed 32 SNPs in *TP63*, providing 24% of the gene coverage. Three of them showed uncorrected significant results in the overall UCB association analysis with a percentage of reproducibility >70% from LASSO. These results warrant an extended UCB study on this region.

Regarding other SNPs in the selected genes, we did not find any strong evidence of association after correcting for multiple testing (permutation test p-value≥0.8 for overall main effects and p-value≥0.3 for subtype effects). The top (uncorrected) significant SNPs were located in *BAK1*, *IGF1R*, *P53AIP1*, *PMAIP1*, *SERPINB5*, and *TP73*. Common variants in these genes have not previously been reported as associated with UCB risk, though an altered expression of *BAK1* and *IGF1R* has been described in bladder tumors.

Many complex diseases, such as UCB, are likely due to the combined effects of multiple loci [32] and most traditional association studies assessing main effects for one SNP at a time are underpowered to detect small effects [33]. Therefore, the implication of common genetic variants may be better assessed by a method that both selects a far-reduced set of potentially associated SNPs and tests for association globally. This has been a challenge due to the high-dimensionality and collinearity

Table 3. Significant SNPs at $\alpha = 0.05$ in the logistic regression main effect models.

		Cases			Controls					Risk of bladder cancer				
GENE	**SNP**	**AA**	**Aa**	**aa**	**AA**	**Aa**	**Aa**	**MAF(a)**	**pHWE**	**OR**	**95% CI**	***p-value***	**MOI**	***Repr. (%)***
BAK1	rs11757379	654	330	42	642	390	54	0.23	0.67	0.86	0.74–0.99	0.047	Add.	33%
IGF1R	rs1058696	968	56	2	998	90	0	0.04	-	0.63	0.44–0.90	0.010	Dom.	81%
IGF1R	rs12591122	758	244	25	824	250	14	0.13	0.34	2.23	1.11–4.51	0.025	Rec.	66%
IGF1R	rs4966015	722	283	21	771	276	41	0.16	0.01	0.44	0.25–0.77	0.004	Rec.	43%
IGF1R	rs702497	633	342	50	645	366	73	0.24	0.04	0.69	0.50–0.94	0.019	Rec.	73%
IGF1R	rs7166348	618	365	45	670	355	62	0.22	0.11	0.67	0.44–1.00	0.050	Rec.	33%
P53AIP1	rs2604235	431	484	109	463	473	149	0.36	0.11	0.74	0.56–0.97	0.029	Rec.	30%
PMAIP1	rs1942919	270	547	207	353	509	224	0.44	0.11	1.27	1.05–1.55	0.015	Dom.	33%
PMAIP1	rs7240884	449	476	100	477	471	138	0.34	0.20	0.75	0.56–0.99	0.047	Rec.	25%
SERPINB5	rs1509476	532	413	82	614	405	69	0.25	0.87	1.20	1.04–1.38	0.012	Add.	14%
SERPINB5	rs1509478	378	490	159	450	493	139	0.36	0.84	1.18	1.04–1.34	0.011	Add.	51%
SERPINB5	rs6567355	466	442	114	552	435	93	0.29	0.60	1.21	1.05–1.38	0.006	Add.	83%
TP63	rs12489753	863	159	5	934	146	7	0.07	0.65	1.31	1.02–1.69	0.035	Dom.	71%
TP63	rs13321831	847	172	8	927	155	6	0.08	1.00	1.36	1.06–1.73	0.014	Dom.	83%
TP63	rs6779677	328	476	224	347	547	194	0.43	0.42	1.29	1.04–1.61	0.022	Rec.	76%
TP73	rs3765731	554	385	86	544	446	96	0.29	0.77	0.85	0.71–1.00	0.050	Dom	71%

MAF(a), minor allele frequency); *pHWE*, *p-value* from the Hardy Weinberg equilibrium test; *MOI*, Mode of Inheritance.
Repr. (%), percentage reproducibility assessing the robustness of each SNP by LASSO.
All models are adjusted for age, gender, region and smoking status.
Odd ratio and 95%CI under the model of inheritance that provided the lowest *p-value*, and percentage reproducibility from LASSO under the additive mode of inheritance.
doi:10.1371/journal.pone.0089952.t003

between SNPs. Nevertheless, penalized techniques can deal with these problems and they are starting to emerge in genetic association studies. Wu et al used penalized logistic regression in a genome-wide association study applied to coeliac disease data and Zhou et al extended this work to the assessment of association for common and rare variants applied to family cancer registry data [34] [35]. In the present study, we applied the LASSO algorithm to account for the combination effects of the SNPs in the TP53 pathway and UCB risk. Under the criteria applied, this method selected one SNP (rs6567355) that showed a non-corrected p-value = 0.006 for the additive mode of inheritance with a percentage of reproducibility = 83%. This is a frequent G> A SNP (MAF = 0.29) located in the intron region of *SERPINB5.* As mentioned before, no evidences of previous association between this SNP and any disease have been reported at present. *SERPINB5* is a tumor suppressor (Table S1 in File S1). The expression levels of this gene has been correlated with those of *DBC1* (Deleted in bladder cancer 1) in UCB specimens, suggesting its involvement in the urokinase-plasminogen pathway [36]. *SERPINB5* would deserve of further exploration in extended studies, as well.

A limitation of our study is the incomplete tagging of the selected genes due to the use of an earlier HapMap release to select tag SNPs, prior to the availability of data from the 1000 genomes project. The median coverage of the 18 genes considered in the pathway is, according to the updated HapMap releases, 44%, ranging from 21% to 86%. Therefore, we cannot rule out completely the implication of common variation in these genes in UCB susceptibility.

For common SNPs (MAF>0.05), our study is powered (90%) to detect ORs≥1.4 at a significance level of 0.05, assuming an additive mode of inheritance. Therefore, the study is not conclusive with OR<1.4. While this study represents one of the largest assessments conducted till present, much larger studies will be required to rule out smaller main effects associated with common variants in the genes of this pathway. This is even more important when subphenotype analyses are considered. We also found no evidence of SNP-SNP interactions (permutation test p-value≥0.3) and SNP-smoking interactions (permutation test p-value≥0.07), although the power was even more limited to detect these. According to the candidate pathway, the studied SNPs were selected as tags; therefore, they were not correlated showing a low LD. This fact, let us overcome a potential limitation affecting the percentage of reproducibility when SNPs are high correlated.

Credit should also be given to this study, not only regarding its large sample size, but also for its prospective nature and disease representativeness, for the homogeneous methods applied to collect information and biosamples by the participating centers, for the integration of different type of information (sociodemographics, epidemiological, genetic, clinical and pathological, and molecular), and for the comprehensive and innovative statistical approaches applied to assess UCB susceptibility associated with a highly candidate pathway.

In conclusion, using a comprehensive analysis accounting different models and different approaches, we found no strong evidence that common variants in the *TP53* pathway are associated with UCB risk. However, specific members of the pathway, *TP63* and *SERPINB5* deserve of further exploration in extended studies. On the other hand, our study suggests that it is unlikely that *TP53* Arg72Pro is implicated in the UCB in white Europeans.

While biological sound, candidate pathway analysis have throw limited acknowledge in the genetic susceptibility field of many diseases. The reasons of this relative poor efficiency may be, among others, the still lack of knowledge of all key components of a given pathway, the introduction of noise by considering many genes/variants without showing association, and the lack of coverage of rare variants not tagged through this approach, in addition to methodological explanations such as an impaired statistical power. Scientists should review whether it is time to dismiss this approach towards a more comprehensive strategy such whole genome/exome sequencing in dissecting the genetic architecture of complex diseases.

Supporting Information

File S1 Combined Supporting Information file containing: Table S1, Location and function of the selected genes. Table S2, Heterogeneity in single nucleotide polymorphism (SNP) risk estimates among bladder cancer subphenotypes defined according to stage and grade in the Spanish Bladder Cancer Study. Table S3, Heterogeneity in single nucleotide polymorphism (SNP) risk estimates among bladder cancer subphenotypes defined by p53 expression in the Spanish Bladder Cancer Study.
(DOCX)

Acknowledgments

We acknowledge the coordinators, field and administrative workers, technicians and patients of the Spanish Bladder Cancer/EPICURO Study.

Author Contributions

Conceived and designed the experiments: SP RLM NR MK DTS FXR NM. Performed the experiments: SJC JL AGN. Analyzed the data: SP RLM MLC ELdM JH LTG EB. Contributed reagents/materials/analysis tools: AN AC AT MM DTS NR MK MGC FXR NM. Wrote the paper: SP RLM NM.

References

1. Ferlay JSH, Bray F, Forman D, Mathers C, Parkin DM (2010) GLOBOCAN 2008 v1.2, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10. Lyon, France: International Agency for Research on Cancer.
2. Samanic C, Kogevinas M, Dosemeci M, Malats N, Real FX, et al. (2006) Smoking and bladder cancer in Spain: effects of tobacco type, timing, environmental tobacco smoke, and gender. Cancer Epidemiol Biomarkers Prev 15: 1348–1354.
3. Malats N (2008) Genetic epidemiology of bladder cancer: scaling up in the identification of low-penetrance genetic markers of bladder cancer risk and progression. Scand J Urol Nephrol Suppl: 131–140.
4. Rothman N, Garcia-Closas M, Chatterjee N, Malats N, Wu X, et al. (2010) A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. Nat Genet 42: 978–984.
5. Luis NM, Lopez-Knowles E, Real FX (2007) Molecular biology of bladder cancer. Clin Transl Oncol 9: 5–12.
6. Wu XR (2005) Urothelial tumorigenesis: a tale of divergent pathways. Nat Rev Cancer 5: 713–725.
7. Guey LT, Garcia-Closas M, Murta-Nascimento C, Lloreta J, Palencia L, et al. (2010) Genetic susceptibility to distinct bladder cancer subphenotypes. Eur Urol 57: 283–292.
8. Real FX, Malats N (2007) Bladder cancer and apoptosis: matters of life and death. Lancet Oncol 8: 91–92.
9. Mitra AP, Hansel DE, Cote RJ (2012) Prognostic value of cell-cycle regulation biomarkers in bladder cancer. Semin Oncol 39: 524–533.
10. Malkin D, Friend SH, Li FP, Strong LC (1997) Germ-line mutations of the p53 tumor-suppressor gene in children and young adults with second malignant neoplasms. N Engl J Med 336: 734.
11. Malkin D, Li FP, Strong LC, Fraumeni JF Jr, Nelson CE, et al. (1990) Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science 250: 1233–1238.

12. Whibley C, Pharoah PD, Hollstein M (2009) p53 polymorphisms: cancer implications. Nat Rev Cancer 9: 95–107.
13. Jiang DK, Ren WH, Yao L, Wang WZ, Peng B, et al. (2010) Meta-analysis of association between TP53 Arg72Pro polymorphism and bladder cancer risk. Urology 76: 765 e761–767.
14. Klug SJ, Ressing M, Koenig J, Abba MC, Agorastos T, et al. (2009) TP53 codon 72 polymorphism and cervical cancer: a pooled analysis of individual data from 49 studies. Lancet Oncol 10: 772–784.
15. Liu KJ, Qi HZ, Yao HL, Lei SL, Lei ZD, et al. (2012) An updated meta-analysis of the p53 codon 72 polymorphism and gastric cancer risk. Mol Biol Rep 39: 8265–8275.
16. Qiao Q, Hu W (2013) The Association Between TP53 Arg72Pro Polymorphism and Lung Cancer Susceptibility: Evidence from 30,038 Subjects. Lung 191: 369–377.
17. Yang ZNS, Zhu H, Wu X, Jia S, Luo Y, et al. (2013) Association of p53 Arg72Pro polymorphism with bladder cancer: a meta-analysis. Gene 512: 408–413.
18. Zhang Z, Wang M, Wu D, Tong N, Tian Y (2009) P53 codon 72 polymorphism contributes to breast cancer risk: a meta-analysis based on 39 case-control studies. Breast Cancer Res Treat 120: 509–517.
19. Epstein JI, Amin MB, Reuter VR, Mostofi FK (1998) The World Health Organization/International Society of Urological Pathology consensus classification of urothelial (transitional cell) neoplasms of the urinary bladder. Bladder Consensus Conference Committee. Am J Surg Pathol 22: 1435–1448.
20. Lorente-Galdos B, Medina I, Morcillo-Suarez C, Heredia T, Carreno-Torres A, et al. (2012) Select your SNPs (SYSNPs): a web tool for automatic and massive selection of SNPs. Int J Data Min Bioinform 6: 324–334.
21. Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81: 1084–1097.
22. Tibshirani R (1996) Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological) 58: 267–288.
23. Friedman J, Hastie T, Thibshirani R (2001) The Elements of statistical learning: Data mining, inference and prediction. Springer Series in Statistics 533 214–216.
24. Efron B (1979) Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics 7: 1–26.
25. Goeman JJ (2010) L1 penalized estimation in the Cox proportional hazards model. Biom J 52: 70–84.
26. Zhou Y, Li N, Zhuang W, Liu GJ, Wu TX, et al. (2007) P53 codon 72 polymorphism and gastric cancer: a meta-analysis of the literature. Int J Cancer 121: 1481–1486.
27. Yan L, Zhang D, Chen C, Mao Y, Xie Y, et al. (2009) TP53 Arg72Pro polymorphism and lung cancer risk: a meta-analysis. Int J Cancer 125: 2903–2911.
28. Naccarati A, Polakova V, Pardini B, Vodickova L, Hemminki K, et al. (2012) Mutations and polymorphisms in TP53 gene–an overview on the role in colorectal cancer. Mutagenesis 27: 211–218.
29. Lin HY, Yang MC, Huang CH, Wu WJ, Yu TJ, et al. (2013) Polymorphisms of TP53 are markers of bladder cancer vulnerability and prognosis. Urol Oncol 31: 1231–1241.
30. Kiemeney LA, Thorlacius S, Sulem P, Geller F, Aben KK, et al. (2008) Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. Nat Genet 40: 1307–1312.
31. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, et al. (2000) Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 343: 78–85.
32. Gibson G (2011) Rare and common variants: twenty arguments. Nat Rev Genet 13: 135–145.
33. Hoh J, Ott J (2003) Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet 4: 701–709.
34. Wu TT, Chen YF, Hastie T, Sobel E, Lange K (2009) Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25: 714–721.
35. Zhou H, Sehl ME, Sinsheimer JS, Lange K (2010) Association screening of common and rare genetic variants by penalized regression. Bioinformatics 26: 2375–2382.
36. Louhelainen JP, Hurst CD, Pitt E, Nishiyama H, Pickett HA, et al. (2006) DBC1 re-expression alters the expression of multiple components of the plasminogen pathway. Oncogene 25: 2409–2419.

FULL PAPER

BJC

British Journal of Cancer (2014) 110, 1088–1100 | doi: 10.1038/bjc.2013.769

Keywords: breast cancer; SNP; FGF receptors; susceptibility; disease subtypes

FGF receptor genes and breast cancer susceptibility: results from the Breast Cancer Association Consortium

D Agarwal[1,2], S Pineda[1], K Michailidou[3], J Herranz[1,4], G Pita[5], L T Moreno[5], M R Alonso[5], J Dennis[3], Q Wang[3], M K Bolla[3], K B Meyer[6], P Menéndez-Rodríguez[7], D Hardisson[8], M Mendiola[9], A González-Neira[5], A Lindblom[10], S Margolin[11], A Swerdlow[12,13], A Ashworth[14], N Orr[14], M Jones[12], K Matsuo[15], H Ito[16], H Iwata[17], N Kondo[17], kConFab Investigators[18], Australian Ovarian Cancer Study Group[18,19], M Hartman[20], M Hui[21], W Y Lim[21], P T-C Iau[22], E Sawyer[23], I Tomlinson[24,25], M Kerin[26], N Miller[26], D Kang[27,28], J-Y Choi[28], S K Park[27,28], D-Y Noh[29], J L Hopper[30], D F Schmidt[30], E Makalic[30], M C Southey[31], S H Teo[32,33], C H Yip[33], K Sivanandan[32], W-T Tay[34], H Brauch[35,36], T Brüning[37], U Hamann[38], The GENICA Network[35,36,37,38,39,40,41], A M Dunning[42], M Shah[42], I L Andrulis[43,44], J A Knight[43,45], G Glendon[46], S Tchatchou[43], M K Schmidt[47], A Broeks[47], E H Rosenberg[47], L J van't Veer[47], P A Fasching[48,49], S P Renner[48], A B Ekici[50], M W Beckmann[48], C-Y Shen[51,52], C-N Hsiung[51], J-C Yu[53], M-F Hou[54,55], W Blot[56], Q Cai[56], A H Wu[57], C-C Tseng[57], D Van Den Berg[57], D O Stram[57], A Cox[58], I W Brock[58], M W R Reed[58], K Muir[59,60], A Lophatananon[60], S Stewart-Brown[60], P Siriwanarangsan[61], W Zheng[62], S Deming-Halverson[62], M J Shrubsole[62], J Long[62], X-O Shu[62], W Lu[63], Y-T Gao[64], B Zhang[62], P Radice[65], P Peterlongo[66], S Manoukian[67], F Mariette[66,68], S Sangrajrang[69], J McKay[70], F J Couch[71], A E Toland[72], TNBCC[73], D Yannoukakos[74], O Fletcher[75], N Johnson[75], I dos Santos Silva[76], J Peto[76], F Marme[77,78], B Burwinkel[77,79], P Guénel[80,81], T Truong[80,81], M Sanchez[80,81], C Mulot[82,83], S E Bojesen[84,85], B G Nordestgaard[84,85], H Flyer[86], H Brenner[87,88], A K Dieffenbach[87,88], V Arndt[87], C Stegmaier[89], A Mannermaa[90,91,92], V Kataja[91,93,94], V-M Kosma[90,91,92], J M Hartikainen[90,91,92], D Lambrechts[95], B T Yesilyurt[95], G Floris[96], K Leunen[96], J Chang-Claude[97], A Rudolph[97], P Seibold[97], D Flesch-Janys[98,99], X Wang[71], J E Olson[100], C Vachon[100], K Purrington[100], G G Giles[30,101], G Severi[30,101], L Baglietto[30,101], C A Haiman[57], B E Henderson[57], F Schumacher[57], L Le Marchand[102], J Simard[103], M Dumont[103], M S Goldberg[104,105], F Labrèche[106], R Winqvist[107], K Pylkäs[107], A Jukkola-Vuorinen[108], M Grip[109], P Devilee[110], R A E M Tollenaar[111], C Seynaeve[112], M García-Closas[13,113,114], S J Chanock[113], J Lissowska[115], J D Figueroa[113], K Czene[116], M Eriksson[116], K Humphreys[116], H Darabi[116], M J Hooning[112], M Kriege[112], J M Collée[117], M Tilanus-Linthorst[118], J Li[119], A Jakubowska[120], J Lubinski[120], K Jaworska-Bieniek[120], K Durda[120], H Nevanlinna[121], T A Muranen[121], K Aittomäki[122], C Blomqvist[123], N Bogdanova[124,125], T Dörk[124], P Hall[116], G Chenevix-Trench[19], D F Easton[3,42], P D P Pharoah[3,42], J I Arias-Perez[126], P Zamora[127], J Benítez[1,5] and R L Milne*[,1,30,101]

Author affiliations can be found at the end of this article

Background: Breast cancer is one of the most common malignancies in women. Genome-wide association studies have identified *FGFR2* as a breast cancer susceptibility gene. Common variation in other fibroblast growth factor (FGF) receptors might also modify risk. We tested this hypothesis by studying genotyped single-nucleotide polymorphisms (SNPs) and imputed SNPs in *FGFR1, FGFR3, FGFR4* and *FGFRL1* in the Breast Cancer Association Consortium.

Methods: Data were combined from 49 studies, including 53 835 cases and 50 156 controls, of which 89 050 (46 450 cases and 42 600 controls) were of European ancestry, 12 893 (6269 cases and 6624 controls) of Asian and 2048 (1116 cases and 932 controls) of African ancestry. Associations with risk of breast cancer, overall and by disease sub-type, were assessed using unconditional logistic regression.

Results: Little evidence of association with breast cancer risk was observed for SNPs in the FGF receptor genes. The strongest evidence in European women was for rs743682 in *FGFR3*; the estimated per-allele odds ratio was 1.05 (95% confidence interval = 1.02–1.09, $P = 0.0020$), which is substantially lower than that observed for SNPs in *FGFR2.*

Conclusion: Our results suggest that common variants in the other FGF receptors are not associated with risk of breast cancer to the degree observed for *FGFR2.*

*Correspondence: RL Milne; E-mail: roger.milne@cancervic.org.au

Received 4 September 2013; revised 8 November 2013; accepted 15 November 2013

Breast cancer is a complex disease, with multiple genetic and environmental factors involved in its etiology. Rare mutations in the DNA repair genes *BRCA1* and *BRCA2* confer a high lifetime risk of breast cancer (Antoniou *et al*, 2003) and are routinely screened for in women with a strong family history of the disease. Studies focused on other DNA repair genes have led to the discovery that rare coding variants in *CHEK2, ATM*, *BRIP1* and *PALB2* (Swift *et al*, 1987; Meijers-Heijboer *et al*, 2002; Seal *et al*, 2006; Rahman *et al*, 2007) are associated with moderately increased breast cancer risk. However, few, if any, candidate-gene- or pathway-based association studies have identified convincing associations with breast cancer risk for common genetic variants (The Breast Cancer Association Consortium, 2006). In contrast, empirical genome-wide association studies (GWAS) have proven to be a successful approach to identify common variants associated with small increases in risk, with more than 70 identified in this way to date (Easton *et al*, 2007; Hunter *et al*, 2007; Stacey *et al*, 2007, 2008; Ahmed *et al*, 2009; Thomas *et al*, 2009; Zheng *et al*, 2009; Antoniou *et al*, 2010; Turnbull *et al*, 2010; Cai *et al*, 2011; Fletcher *et al*, 2011; Haiman *et al*, 2011; Ghoussaini *et al*, 2012; Siddiq *et al*, 2012; Bojesen *et al*, 2013; Garcia-Closas *et al*, 2013; Michailidou *et al*, 2013). For the great majority of these associations, the causal variant(s), and even the causal gene, are unknown; thus, the identification of novel candidate genetic susceptibility pathways through this approach is not straightforward.

An intronic variant in the *FGFR2* gene was one of the first single-nucleotide polymorphisms (SNPs) identified by GWAS as tagging a breast cancer susceptibility locus (Easton *et al*, 2007; Hunter *et al*, 2007). It is now well-established that the minor allele of this SNP is associated with increased risk of breast cancer, particularly estrogen receptor (ER)-positive disease (Garcia-Closas *et al*, 2008). Fine-mapping of the region has suggested that at least one causal variant is located in intron 2 of *FGFR2* (Easton *et al*, 2007; Udler *et al*, 2009), and functional studies have proposed that rs2981578 affects *FGFR2* expression (Meyer *et al*, 2008; Udler *et al*, 2009; Huijts *et al*, 2011). These findings strongly suggest that *FGFR2* is a breast cancer susceptibility gene.

FGFR2 is a fibroblast growth factor (FGF) receptor gene; the amino-acid sequence of the protein it encodes is highly conserved across all FGF receptors. The other FGF receptor genes and other genes acting downstream of them in the FGF pathway may also be implicated in the development of breast cancer, although associations with disease risk have not been assessed comprehensively by a study with adequate sample size to detect odds ratios (ORs) of the magnitude observed for SNPs in *FGFR2*.

We hypothesised that common variants in other genes in the FGF pathway, and in the other FGF receptor genes in particular, might also confer increased breast cancer risk. The primary aim of our investigation was to comprehensively assess associations between breast cancer risk and common variation in the FGF receptor genes *FGFR1, FGFR3, FGFR4* and *FGFRL1* by genotyping selected tag-SNPs in the Breast Cancer Association Consortium (BCAC). A secondary objective was to assess common variants in other genes in the FGF pathway based on a two-stage design.

MATERIALS AND METHODS

Participants. Study participants were women from 49 studies participating in BCAC: 38 from populations of predominantly European ancestry, 9 of Asian women and 2 of African–American women (Table 1 and Supplementary Table 1). The majority were population-based or hospital-based case–control studies, but some studies selected subjects based on age or oversampled for cases with a family history or bilateral disease. Cases and controls from the CNIO-BCS were also studied in a previous assessment of selected genes in the FGF pathway. All study participants gave informed consent and each study was approved by the corresponding local ethics committee.

Gene and SNP selection. Ingenuity Pathways Analysis and selected publications (Eswarakumar *et al*, 2005; Presta *et al*, 2005; Chen & Forough, 2006; Schwertfeger, 2009) were used to identify genes reported to be involved downstream of the *FGF* genes in the FGF pathway, particularly those related to angiogenesis. A total of 39 genes, including the FGF receptors *FGFR1* (located at 8p11.22), *FGFR2* (10q26.13), *FGFR3* (4p16.3), *FGFR4* (5q35.2) and *FGFRL1* (4p16.3), was selected for tagging. Single-nucleotide polymorphisms with minor allele frequency (MAF) $>5\%$ in the coding and non-coding regions, and within 5 kb upstream and 5 kb downstream of each gene, were identified using HapMap CEU genotype data and dbSNP 128 as reference. The minimum number of tag-SNPs were then selected among all identified SNP using Haploview (Barrett *et al*, 2005) based on the following criteria: $r^2>0.8$ and Illumina assay score >0.60. A total of 384 SNPs tagging 39 genes was genotyped in the CNIO-BCS, 324 of which were successfully genotyped (Supplementary Table 2). The 31 SNPs tagging genes *FGFR1, FGFR3, FGFR4* and *FGFRL1* were all genotyped in BCAC, along with a further 26 of the 324 tag-SNPs. The latter group comprised SNPs selected based on evidence of association with breast cancer under a log-additive model in the Stage 1 CNIO-BCS. Single-nucleotide polymorphisms in *FGFR2* were not considered, as all were included as part of a separate fine-mapping study (Meyer *et al*, submitted). Results from Stage 1 are summarised in Supplementary Table 2.

Genotyping. Genotyping of the 57 SNPs in the BCAC samples was conducted using a custom Illumina Infinium array (iCOGS) in four centers, as part of a multi-consortia collaboration (the Collaborative Oncological Gene–Environment Study, COGS) as described previously (Michailidou *et al*, 2013). Genotypes were called using Illumina's proprietary GenCall algorithm.

For the genotyping of the 384 SNPs in the Stage 1 CNIO-BCS, genomic DNA was isolated from peripheral blood lymphocytes using automatic DNA extraction (MagNA Pure, Roche Diagnostics, Indianapolis, IN, USA) according to the manufacturer's recommended protocols. This DNA was quantified using Picogreen (Invitrogen, Life Technologies, Grand Island, NY, USA) and for each sample a final quantity of 250 ng was extracted and used for GoldenGate genotyping with VeraCode Technology (Illumina Inc., San Diego, CA, USA). Samples were arranged on 25 96-well plates containing one negative control and at least one study sample in duplicate. Three Centre d'Etude du Polymorphisme Humain (CEPH) trios were used as internal intra- and inter-plate duplicates and to check for Mendelian segregation errors. DNA was extracted, quantified, plated and genotyped at the Spanish National Genotyping Centre (CeGen), Madrid, Spain. All genotypes were determined for each SNP and each plate using manual clustering. Single-nucleotide polymorphisms with call rate $<90\%$ were excluded, as were samples with no-calls for more than 20% of included SNPs.

Statistical methods. For each SNP, we estimated ORs and 95% confidence intervals (CIs) using unconditional logistic regression. For the analysis of BCAC data, we considered per-allele and co-dominant models using common-allele homozygotes as reference and including study and ethnicity-specific principal components as covariates, as previously described (Michailidou *et al*, 2013). Departure from the Hardy–Weinberg equilibrium (HWE) was tested for in controls from individual studies using the *genhwi* module in STATA 11.2 (College Station, TX, USA). A study-stratified χ^2 test (1df) was applied across studies (Haldane, 1954; Robertson & Hill, 1984). Between-study heterogeneity in ORs was assessed for each of the three broad racial groups using the *metan* command in STATA to meta-analyse study-specific per-allele log-OR estimates and generate I^2 statistics; values greater than 50% were considered notable (Higgins & Thompson, 2002). Odds ratios specific to disease subtypes defined by ER, PR and HER2 status (positive and negative) were estimated separately for each ethnic subgroup using polytomous logistic regression with control status as the reference outcome. Differences in ORs by disease subtypes were assessed using a likelihood ratio test (LRT). All statistical tests were two-sided.

The effective number of independent SNPs (V_{effLi}) was estimated using the method described by Li & Ji (2005). This method was applied via the web-interface matSpDlite (http://gump.qimr.edu.au/general/daleN/matSpDlite/), based on the observed correlations between SNPs (Nyholt, 2004). V_{effLi} was then used to calculate a Bonferroni-corrected significance threshold (α^*). Power calculations were carried out using Quanto v1.2.4 (http://hydra.usc.edu/gxe/).

Single-nucleotide polymorphism imputation. The genotypes of untyped SNPs were imputed based on data from the March 2012 release of the 1000 genomes project using IMPUTE v2.2. These were converted to allele doses using the *impute2mach* function in the *GenABEL* library in R (Aulchenko *et al*, 2007) and analysed under a per-allele model. Imputed SNPs with an estimated MAF $<5\%$ were excluded, as were SNPs with an imputation $r^2<80\%$.

RESULTS

All SNPs in the present analysis had overall call rates $>95\%$. Very strong evidence of departure from HWE was observed for rs34869253 for one study (pKarma, $P=3.3\times10^{-21}$), which was excluded from the subsequent analyses of that SNP. After quality control, there were data available for 53 835 cases and 50 156 controls from BCAC, including 89 050 European women (46 450 cases and 42 600 controls), 12 893 Asian (6269 cases and 6624 controls) and 2048 African–American women (1116 cases and 932 controls) (Table 1).

Results from the analysis of the 31 tag-SNPs in *FGFR* genes for white Europeans are summarised in Table 2. No strong evidence of association was observed, although one SNP (rs743682) in *FGFR3* (MAF = 9%) was marginally significant after correction for multiple testing based on a V_{effLi} of 23 (per-allele OR = 1.05, 95%CI = 1.02–1.09, $P=0.0020$, $\alpha^*=0.0022$). All SNPs with an associated *P*-value <0.05 were intronic, with the exception of rs1966265, which is a missense variant in *FGFR4*. However, PolyPhen (http://genetics.bwh.harvard.edu/pph2/) predicts this amino acid change to be benign, with a score of 0.000. On the basis of ENCODE data, no SNP with an associated *P*-value <0.05 was located in a region involved or predicted to be involved in epigenetic regulation, nor at, or within 2 kb of, a CpG island. For European women, we did not observe any evidence of between-study heterogeneity for any SNPs ($I^2\leqslant19\%$; $P\geqslant0.15$) and little evidence of differential associations by disease subtypes defined by ER ($P\geqslant0.036$), PR ($P\geqslant0.084$) or HER2 status ($P\geqslant0.019$).

We similarly observed little evidence of association with overall breast cancer risk in Asian and African–American women (Supplementary Tables 3 and 4, respectively). Nevertheless, a consistent result was observed for Europeans and Asians for rs1966265 in *FGFR4*. The estimated OR per risk (G) allele was 1.03 (95%CI = 1.01–1.05; $P=0.0060$) for European women and 1.08 (95%CI = 1.03–1.14; $P=0.0036$) for Asian women. There was no evidence of heterogeneity by race for any of the 31 SNPs in FGF receptors ($I^2=18\%$; $P=0.14$).

The SNPs genotyped were estimated to capture a variable proportion of the common variation in the four genes considered, as described in the 1000 genomes project; at $r^2\geqslant0.80$, this coverage was 75% for *FGFR1*, 77% for *FGFR3*, 66% for *FGFR4* and 17% for *FGFRL1*. This coverage was dramatically improved with the inclusion of imputed common SNPs (with imputation $r^2>0.80$) to 95%, 93%, 97% and 84% for *FGFR1, FGFR3, FGFR4* and *FGFRL1,* respectively. No stronger evidence of association was observed for any imputed SNPs (Supplementary Tables 5–8).

Finally, we observed little evidence of association for any of the 26 SNPs in other genes in the FGF pathway, selected based on results from Stage 1 (Supplementary Table 9). The results were consistent across the three ethnic groups considered and for disease subtypes defined by ER, PR and HER2 expression.

It is noteworthy that strong association signals were observed in the Stage 1 Spanish study for selected tag-SNPs rs10736303 (MAF = 0.49; per-allele OR = 1.37, 95% CI = 1.21–1.55, $P=2.8\times10^{-7}$), and rs2981582 (MAF = 0.40; per-allele OR = 1.35, 95% CI = 1.19–1.53, $P=8.3\times10^{-7}$), both in *FGFR2.*

DISCUSSION

In this multicentre case–control study, we comprehensively assessed common variation in the FGF receptor genes *FGFR1, FGFR3, FGFR4* and *FGFRL1* in 53 835 cases and 50 156 controls and found little evidence of association with risk of breast cancer. This is the largest study we know of assessing a family of genes via a candidate approach based on the findings from GWAS.

A non-trivial issue in analyses of this kind is the establishment of a statistical significance threshold that adequately controls the proportion of false-positive findings. As permutation-testing was not feasible due to the sample size and number of dummy variables required to adjust for study, we dealt with the issue of non-independence of multiple tests by estimating that the 31 tag-SNPs represented an effective number of 23 independent variables, and applying a Bonferroni correction accordingly. The association of one SNP (rs743682) in *FGFR3* for European women was found to be statistically significant on this basis. However, the *P*-value threshold applied is somewhat questionable in the context of the total of more than 70 000 SNPs nominated for genotyping by BCAC and the total 210 000 genotyped on the iCOGS array. Thus, the current result is far from genome-wide statistical significance and certainly requires independent replication. In any case, the per-allele ORs for *FGFR3*_rs743682 (1.05, 95% CI = 1.02–1.09) and *FGFR4*_rs1966265 (1.03, 95% CI = 1.01–1.05) appear to be substantially lower than that for rs2981582 in *FGFR2* (1.26, 95% CI = 1.23–1.30) (Easton *et al*, 2007).

We estimated that for common SNPs (MAF >0.05) associated with overall breast cancer risk in European women, we had greater than 99% power to detect at genome-wide statistical significance ($P<5\times10^{-8}$) a per-allele OR as low as 1.23 (the lower 95% confidence limit for the OR for *FGFR2*_rs2981582). For a per-allele OR as low as 1.05 and for SNPs with MAF of 0.10, 0.20 and 0.30, the estimated power was 1%, 10% and 24%, respectively. That is, our study provides strong evidence that common variation in *FGFR1, FGFR3, FGFR4* and *FGFRL1* are not associated with breast cancer risk to the degree observed for SNPs in *FGFR2*, although associations of smaller magnitude may exist.

The hypothesis underlying our study was based on the identification of a functional SNP in intron 2 of *FGFR2* associated with breast cancer susceptibility (Easton *et al*, 2007; Meyer *et al*, 2008; Udler *et al*, 2009; Huijts *et al*, 2011). A recent study has subsequently identified three independent risk signals within *FGFR2*, and uncovered likely causal variants and functional mechanisms behind them (Meyer *et al*, 2013). Although an association between these SNPs and expression of FGFR2 has not been established, these results provide strong

Table 1. Number of cases and controls included, by study

Study	Country	Controls	Cases	ER+	ER−
White European women					
Australian Breast Cancer Family Study[a] (ABCFS)	Australia	551	790	456	261
Amsterdam Breast Cancer Study (ABCS)	Netherlands	1429	1325	420	153
Bavarian Breast Cancer Cases and Controls (BBCC)	Germany	458	564	460	83
British Breast Cancer Study (BBCS)	UK	1397	1554	507	114
Breast Cancer In Galway Genetic Study (BIGGS)	Ireland	719	836	495	154
Breast Cancer Study of the University Clinic Heidelberg (BSUCH)	Germany	954	852	499	154
CECILE Breast Cancer Study (CECILE)	France	999	1019	797	144
Copenhagen General Population Study (CGPS)	Denmark	4086	2901	1919	357
Spanish National Cancer Centre Breast Cancer Study (CNIO-BCS)	Spain	876	902	242	88
California Teachers Study (CTS)	USA	71	68	0	17
ESTHER Breast Cancer Study (ESTHER)	Germany	502	478	304	98
Gene–Environment Interaction and Breast Cancer in Germany (GENICA)	Germany	427	465	328	119
Helsinki Breast Cancer Study (HEBCS)	Finland	1234	1664	1295	237
Hannover-Minsk Breast Cancer Study (HMBCS)	Belarus	130	690	37	0
Karolinska Breast Cancer Study (KARBAC)	Sweden	662	722	338	63
Kuopio Breast Cancer Project (KBCP)	Finland	251	445	304	97
kConFab/Australian Ovarian Cancer Study (kConFab/AOCS)	Australia	897	613	162	59
Leuven Multidisciplinary Breast Centre (LMBC)	Belgium	1388	2671	2071	379
Mammary Carcinoma Risk Factor Investigation (MARIE)	Germany	1778	1818	1349	399
Milan Breast Cancer Study Group (MBCSG)	Italy	400	488	149	42
Mayo Clinic Breast Cancer Study (MCBCS)	USA	1931	1862	1486	295
Melbourne Collaborative Cohort Study (MCCS)	Australia	511	614	352	119
Multi-ethnic Cohort (MEC)	USA	741	731	415	87
Montreal Gene–Environment Breast Cancer Study (MTLGEBCS)	Canada	436	489	421	64
Norwegian Breast Cancer Study (NBCS)	Norway	70	22	0	22
Oulu Breast Cancer Study (OBCS)	Finland	414	507	407	100
Ontario Familial Breast Cancer Registry[b] (OFBCR)	Canada	511	1175	630	268
Leiden University Medical Centre Breast Cancer Study (ORIGO)	Netherlands	327	357	211	70
NCI Polish Breast Cancer Study (PBCS)	Poland	424	519	519	0
Karolinska Mammography Project for Risk Prediction of Breast Cancer (pKARMA)	Sweden	5537	5434	3672	702
Rotterdam Breast Cancer Study (RBCS)	Netherlands	699	664	368	131
Singapore and Sweden Breast Cancer Study (SASBAC)	Sweden	1378	1163	663	144
Sheffield Breast Cancer Study (SBCS)	UK	848	843	377	105
Studies of Epidemiology and Risk factors in Cancer Heredity (SEARCH)	UK	8069	9347	5160	1181
Städtisches Klinikum Karlsruhe Deutsches Krebsforschungszentrum Study (SKKDKFZS)	Germany	29	136	0	136
IHCC-Szczecin Breast Cancer Study (SZBCS)	Poland	315	365	165	60
Triple Negative Breast Cancer Consortium Study (TNBCC)	Various	542	881	0	881
UK Breakthrough Generations Study (UKBGS)	UK	470	476	96	22
Asian women					
Asian Cancer Project (ACP)	Thailand	636	423	92	53
Hospital-based Epidemiologic Research Program at Aichi Cancer Center (HERPACC)	Japan	1376	694	395	139
Los Angeles County Asian-American Breast Cancer Case–Control (LAABC)	USA	990	812	528	138
Malaysian Breast Cancer Genetic Study (MYBRCA)	Malaysia	610	770	422	291
Shanghai Breast Cancer Genetic Study (SBCGS)	China	892	848	510	276
Seoul Breast Cancer Study (SEBCS)	South Korea	1129	1162	657	375
Singapore Breast Cancer Cohort (SGBCC)	Singapore	502	533	272	108
IARC-Thai Breast Cancer (TBCS)	Thailand	253	138	26	26
Taiwanese Breast Cancer Study (TWBCS)	Taiwan	236	889	460	204
African					
Southern Community Cohort Study (SCCS)	USA	680	679	0	0
Nashville Breast Health Study (NBHS)	USA	252	437	199	222
Total		50156	53835	30635	9120

Abbreviations: ER− = estrogen receptor-negative cases; ER+ = estrogen receptor-positive cases.

[a] Australian site of the Breast Cancer Family Registry.

[b] Ontario site of the Breast Cancer Family Registry.

evidence that *FGFR2* is the target gene, and it therefore seems plausible that other FGF receptors or genes acting in the FGF pathway might also be implicated in breast cancer risk. However, we find little evidence that this is the case for the receptors, at least not to the extent observed for common variants in *FGFR2*. Admittedly, the degree to which common variation in the FGF

Table 2. Summary results for SNPs in FGF receptor genes for white European women

			OR (95%CI)				OR (95%CI)		
SNP	**Alleles**	**MAF**	**Aa**	**aa**	**per-a-allele**	***P***	**ER−**	**ER+**	***P-het***
FGFR1									
rs10958704	AG	0.40	0.98 (0.95–1.01)	0.98 (0.94–1.02)	0.99 (0.97–1.01)	0.18	0.99 (0.96–1.03)	0.99 (0.97–1.02)	0.91
rs17182141	AG	0.06	1.05 (1.00–1.09)	0.95 (0.75–1.22)	1.04 (1.00–1.08)	0.057	1.08 (1.00–1.17)	1.04 (0.99–1.09)	0.30
rs2288696	GA	0.21	1.02 (0.99–1.05)	1.07 (1.00–1.14)	1.03 (1.00–1.05)	0.023	1.05 (1.01–1.10)	1.03 (1.00–1.06)	0.35
rs2411256	GA	0.24	1.02 (0.99–1.05)	1.01 (0.95–1.07)	1.01 (0.99–1.03)	0.36	1.00 (0.95–1.04)	1.01 (0.99–1.04)	0.44
rs2978076	GA	0.08	0.99 (0.96–1.03)	1.22 (1.04–1.44)	1.01 (0.98–1.05)	0.53	0.99 (0.92–1.06)	1.02 (0.98–1.06)	0.37
rs2978083	GA	0.05	0.99 (0.96–1.03)	1.22 (1.04–1.44)	1.01 (0.98–1.05)	0.53	0.97 (0.89–1.06)	1.03 (0.97–1.08)	0.27
rs3758102	GA	0.26	1.01 (0.98–1.04)	1.02 (0.96–1.07)	1.01 (0.99–1.03)	0.35	1.01 (0.97–1.05)	1.01 (0.98–1.04)	0.95
rs3925	GA	0.23	1.01 (0.98–1.04)	1.00 (0.95–1.07)	1.01 (0.99–1.03)	0.51	0.99 (0.95–1.04)	1.01 (0.99–1.04)	0.39
rs4733930	GA	0.42	1.00 (0.97–1.03)	1.04 (1.00–1.08)	1.02 (1.00–1.04)	0.11	1.03 (0.99–1.07)	1.02 (1.00–1.04)	0.67
rs4733946	CA	0.08	1.00 (0.97–1.03)	1.04 (1.00–1.08)	1.02 (1.00–1.04)	0.11	1.01 (0.95–1.08)	1.04 (1.00–1.09)	0.39
rs6474354	GA	0.21	0.98 (0.95–1.01)	0.99 (0.92–1.05)	0.98 (0.96–1.01)	0.18	0.96 (0.92–1.01)	0.98 (0.96–1.01)	0.37
rs6996321	GA	0.39	1.01 (0.98–1.04)	1.00 (0.96–1.04)	1.00 (0.98–1.02)	0.95	1.00 (0.97–1.04)	0.99 (0.97–1.02)	0.54
rs6983315	GA	0.44	1.01 (0.97–1.04)	0.98 (0.94–1.02)	0.99 (0.97–1.01)	0.39	0.97 (0.93–1.00)	0.99 (0.97–1.02)	0.13
rs7012413	GA	0.30	1.00 (0.97–1.02)	0.99 (0.95–1.04)	1.00 (0.98–1.02)	0.69	1.00 (0.97–1.04)	1.00 (0.98–1.02)	0.82
FGFR3									
rs12502543	GA	0.10	1.04 (1.01–1.08)	1.10 (0.96–1.25)	1.04 (1.01–1.08)	0.0076	0.99 (0.93–1.05)	1.06 (1.02–1.10)	0.036
rs2234909	AG	0.14	0.99 (0.95–1.02)	0.97 (0.88–1.07)	0.99 (0.96–1.01)	0.29	0.99 (0.94–1.04)	0.98 (0.95–1.02)	0.77
rs3135848	AG	0.28	1.02 (0.99–1.04)	1.02 (0.96–1.07)	1.01 (0.99–1.03)	0.31	1.00 (0.96–1.04)	1.01 (0.99–1.04)	0.55
rs743682	GA	0.09	1.05 (1.01–1.09)	1.16 (1.00–1.34)	1.05 (1.02–1.09)	0.0020	1.01 (0.95–1.08)	1.06 (1.02–1.10)	0.24
rs746779	GA	0.18	0.99 (0.96–1.02)	0.98 (0.90–1.06)	0.99 (0.96–1.01)	0.29	1.00 (0.95–1.05)	0.98 (0.95–1.01)	0.48
FGFR4									
rs1076891	GA	0.06	1.03 (0.99–1.08)	0.99 (0.81–1.22)	1.03 (0.99–1.07)	0.14	1.06 (0.98–1.14)	1.01 (0.97–1.06)	0.25
rs1966265	GA	0.23	0.97 (0.94–1.00)	0.93 (0.88–0.99)	0.97 (0.95–0.99)	0.0060	0.98 (0.94–1.03)	0.97 (0.95–1.00)	0.54
rs2456173	GA	0.21	1.00 (0.97–1.03)	0.99 (0.92–1.05)	0.99 (0.97–1.02)	0.66	0.98 (0.94–1.02)	1.00 (0.98–1.03)	0.34
rs376618	AG	0.24	1.00 (0.97–1.03)	0.96 (0.91–1.02)	0.99 (0.97–1.01)	0.33	0.97 (0.93–1.01)	0.99 (0.97–1.02)	0.29
rs641101	GA	0.31	1.01 (0.98–1.04)	0.99 (0.94–1.03)	1.00 (0.98–1.02)	0.98	0.99 (0.95–1.03)	1.00 (0.98–1.02)	0.56
rs6556301	CA	0.36	0.99 (0.97–1.02)	0.96 (0.92–1.00)	0.98 (0.97–1.00)	0.13	0.99 (0.95–1.02)	0.98 (0.96–1.01)	0.84
FGFRL1									
rs34869253	AG	0.43	1.00 (0.97–1.04)	1.00 (0.96–1.04)	1.00 (0.98–1.02)	0.96	0.98 (0.94–1.01)	0.99 (0.97–1.01)	0.52
rs3755955	GA	0.16	1.00 (0.97–1.03)	1.02 (0.94–1.11)	1.00 (0.98–1.03)	0.82	1.00 (0.95–1.05)	1.00 (0.97–1.03)	0.83
rs4505759	GA	0.30	0.99 (0.96–1.02)	0.98 (0.93–1.03)	0.99 (0.97–1.00)	0.38	1.00 (0.96–1.04)	0.99 (0.97–1.02)	0.78
rs4647932	GA	0.06	1.04 (0.99–1.08)	0.98 (0.80–1.20)	1.03 (0.99–1.07)	0.14	1.06 (0.98–1.14)	1.02 (0.97–1.06)	0.31
rs6855233	AG	0.29	0.99 (0.97–1.02)	1.03 (0.98–1.08)	1.01 (0.98–1.03)	0.62	0.98 (0.94–1.02)	1.00 (0.98–1.03)	0.31
rs748651	AG	0.48	1.00 (0.97–1.03)	1.02 (0.98–1.06)	1.01 (0.99–1.03)	0.31	1.03 (0.99–1.07)	1.01 (0.98–1.03)	0.22

Abbreviations: SNP = single-nucleotide polymorphism; FGF = fibroblast growth factor; OR = odds ratio where A is the common allele, a is the rare allele and both Aa and aa are compared with AA genotypes; CI = confidence interval; MAF = minor allele frequency; *P* = *P*-value for the per-a-allele model; ER− = results (per a-allele) for risk of estrogen receptor-negative disease; ER+ = results (per a-allele) for risk of estrogen receptor-positive disease; *P-het* = *P*-value for heterogeneity by disease sub-type defined by estrogen receptor status.

receptor genes was tagged by the genotyped SNPs was good for *FGFR1, FGFR3* and *FGFR4* and poor for *FGFRL1*, but substantial improvement was afforded by imputation. Nevertheless, it is possible that common variation not captured by the genotyped or imputed SNPs may be associated with breast cancer risk. It is also possible that these genes may be implicated in disease susceptibility via regulatory mechanisms involving variants outside the chromosomal boundaries defined for each gene considered. That said, few studies have assessed common variation in candidate genes to this extent, in terms of both gene coverage and sample size.

The power of our study was much lower for Asian and African–American women; however, our primary focus on European women is consistent with our hypothesis, based on the previous finding in *FGFR2* in this population. Our study was also limited by the power and gene coverage of the stage 1 component which assessed tag-SNPs in the selected genes of the FGF pathway. Therefore, no conclusions can be drawn about the potential implication of common variation in these genes in breast cancer susceptibility. Nevertheless, we checked the chromosomal locations of the 76 established risk-associated loci (http://www.nature.com/icogs/primer/shared-susceptibility-loci-for-breast-prostate-and-ovarian-cancers/) and found that none were located within 10 kb of any of the 39 genes considered, with the exception of the *FGFR2* locus.

In conclusion, in this, possibly the largest candidate-gene association study carried out to date, we have observed little evidence of association between common variation in the *FGFR1, FGFR3, FGFR4* and *FGFRL1* genes and risk of breast cancer. Our results suggest that common variants in these FGF receptors are not associated with risk of breast cancer to the degree observed for *FGFR2*.

ACKNOWLEDGEMENTS

We thank all the individuals who took part in these studies and all the researchers, study staff, clinicians and other health-care providers, technicians and administrative staff who have enabled this work to be carried out. In particular, we thank: Andrew Berchuck (OCAC), Rosalind A Eeles, Ali Amin Al Olama, Zsofia Kote-Jarai, Sara Benlloch (PRACTICAL), Antonis Antoniou, Lesley McGuffog, Ken Offit (CIMBA), Andrew Lee, and Ed Dicks, Craig Luccarini and the staff of the Centre for Genetic Epidemiology Laboratory, Daniel C Tessier, Francois Bacot, Daniel Vincent, Sylvie LaBoissière, Frederic Robidoux and the staff of the McGill University and Génome Québec Innovation Centre, Sune F Nielsen, and the staff of the Copenhagen DNA laboratory, Julie M Cunningham, Sharon A Windebank, Christopher A Hilker, Jeffrey Meyer and the staff of Mayo Clinic Genotyping Core Facility, Maggie Angelakos, Judi Maskiell, Sten Cornelissen, Ellen van der Schoot, Emiel Rutgers, Senno Verhoef, Frans Hogervorst, the Thai Ministry of Public Health (MOPH), Dr Prat Boonyawongviroj (former Permanent Secretary of MOPH), Dr Pornthep Siriwanarungsan (Department Director-General of Disease Control), Alexander Hein, Michael Schrauder, Matthias Rübner, Sonja Oeser, Silke Landrith, Eileen Williams, Elaine Ryder-Mills, Kara Sargus, Niall McInerney, Gabrielle Colleran, Andrew Rowan, Angela Jones, Christ of Sohn, Andeas Schneeweiß, Peter Bugert, the Danish Breast Cancer Group, Ana Osorio, Alicia Barroso, Victoria Fernández, Maika González-Neira, Núria Malats, Francisco Real, Belén Herráez, Núria Álvarez, Esther Díaz, María Miguel-Martin, the CTS Steering Committee (including: Leslie Bernstein, James Lacey, Sophia Wang, Huiyan Ma, Yani Lu and Jessica Clague DeHart at the Beckman Research Institute of the City of Hope; Dennis Deapen, Rich Pinder, Eunjung Lee and Fred Schumacher at the University of Southern California; Pam Horn-Ross, Peggy Reynolds and David Nelson at the Cancer Prevention Institute of California; and Hannah Park at the University of California Irvine), Hartwig Ziegler, Sonja Wolf, Volker Hermann, The GENICA network (Dr Margarete Fischer-Bosch Institute of Clinical Pharmacology, Stuttgart, and University of Tübingen, Germany; (HB, Wing-Yee Lo, Christina Justenhoven), Department of Internal Medicine, Evangelische Kliniken Bonn GmbH, Johanniter Krankenhaus, Bonn, Germany (Yon-Dschun Ko, Christian Baisch), Institute of Pathology, University of Bonn, Germany (Hans-Peter Fischer), Molecular Genetics of Breast Cancer, Deutsches Krebsforschungszentrum (DKFZ) Heidelberg, Germany (UH), Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Germany (TB, Beate Pesch, Sylvia Rabstein, Anne Lotz), Institute of Occupational Medicine and Maritime Medicine, University Medical Center Hamburg-Eppendorf, Germany (Volker Harth)), Tuomas Heikkinen, Irja Erkkilä, Kirsimari Aaltonen, Karl von Smitten, Natalia Antonenkova, Peter Hillemanns, Hans Christiansen, Eija Myöhänen, Helena Kemiläinen, Heather Thorne, Eveline Niedermayr, the AOCS Management Group (D Bowtell, G Chenevix-Trench, A deFazio, D Gertig, A Green, P Webb), the ACS Management Group (A Green, P Parsons, N Hayward, P Webb and D Whiteman), the LAABC data collection team, especially Annie Fung and June Yashiki, Gilian Peuteman, Dominiek Smeets, Thomas Van Brussel, Kathleen Corthouts, Nadia Obi, Judith Heinz, Sabine Behrens, Ursula Eilber, Muhabbet Celik, Til Olchers, Bernad Peissel, Giulietta Scuvera, Daniela Zaffaroni, Bernardo Bonanni, Monica Barile, Irene Feroce, the personnel of the Cogentech Cancer Genetic Test Laboratory, The Mayo Clinic Breast Cancer Patient Registry, Martine Tranchant, Marie-France Valois, Annie Turgeon, Lea Heguy, Phuah Sze Yee, Peter Kang, Kang In Nee, Shivaani Mariapun, Yoon Sook-Yee, Daphne Lee, Teh Yew Ching, Nur Aishah Mohd Taib, Meeri Otsukka, Kari Mononen, Teresa Selander, Nayana Weerasooriya, OFBCR staff, E Krol-Warmerdam, J Molenaar, J Blom, Louise Brinton, Neonila Szeszenia-Dabrowska, Beata Peplonska, Witold Zatonski, Pei Chao, Michael Stagner, Petra Bos, Jannet Blom, Ellen Crepin, Anja Nieuwlaat, Annette Heemskerk, the Erasmus MC Family Cancer Clinic, Sue Higham, Simon Cross, Helen Cramp, Dan Connley, The Eastern Cancer Registration and Information Centre, the SEARCH and EPIC teams, Craig Luccarini, Don Conroy, Caroline Baynes, Kimberley Chua, the Ohio State University Human Genetics Sample Bank and Robert Pilarski. Data on SCCS cancer cases used in this publication were provided by the: Alabama Statewide Cancer Registry; Kentucky Cancer Registry, Lexington, KY; Tennessee Department of Health, Office of Cancer Surveillance; Florida Cancer Data System; North Carolina Central Cancer Registry, North Carolina Division of Public Health; Georgia Comprehensive Cancer Registry; Louisiana Tumor Registry; Mississippi Cancer Registry; South Carolina Central Cancer Registry; Virginia Department of Health, Virginia Cancer Registry; Arkansas Department of Health, Cancer Registry. BCAC is funded by Cancer Research UK (C1287/A10118, C1287/A12014) and by the European Community's Seventh Framework Programme under grant agreement number 223175 (HEALTH-F2-2009-223175) (COGS). Meetings of the BCAC have been funded by the European Union COST programme (BM0606). Genotyping of the iCOGS array was funded by the European Union (HEALTH-F2-2009-223175), Cancer Research UK (C1287/A10710), the Canadian Institutes of Health Research for the 'CIHR Team in Familial Risks of Breast Cancer' programme, and the Ministry of Economic Development, Innovation and Export Trade of Quebec (PSR-SIIRI-701). Additional support for the iCOGS infrastructure was provided by the National Institutes of Health (CA128978) and Post-Cancer GWAS initiative (1U19 CA148537, 1U19 CA148065 and 1U19 CA148112 – the GAME-ON initiative), the Department of Defence (W81XWH-10-1-0341), Komen Foundation for the Cure, the Breast Cancer Research Foundation, and the Ovarian Cancer Research Fund. The ABCFS and OFBCR work was supported by grant UM1 CA164920 from the National Cancer Institute (USA). The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR) nor does mention of trade names, commercial products or organizations imply endorsement by the US Government or the BCFR. The ABCFS was also supported by the National Health and Medical Research Council of Australia, the New South Wales Cancer Council, the Victorian Health Promotion Foundation (Australia) and the Victorian Breast Cancer Research Consortium. JLH is a National Health and Medical Research Council (NHMRC) Senior Principal Research Fellow and MCS is a NHMRC Senior Research Fellow. The OFBCR work was also supported by the Canadian Institutes of Health Research 'CIHR Team in Familial Risks of Breast Cancer' program. The ABCS was funded by the Dutch Cancer Society Grant no. NKI2007-3839 and NKI2009-4363. The ACP study is funded by the Breast Cancer Research Trust, UK. The work of the BBCC was partly funded by ELAN-Programme of the University Hospital of Erlangen. The BBCS is funded by Cancer Research UK and Breakthrough Breast Cancer and acknowledges NHS funding to the NIHR Biomedical Research Centre and the National Cancer Research Network (NCRN). ES is supported by NIHR Comprehensive Biomedical Research Centre, Guy's & St. Thomas' NHS Foundation Trust in partnership with King's College London, United Kingdom. IT is supported by the Oxford Biomedical Research Centre. The BSUCH study was supported by the Dietmar-Hopp Foundation, the Helmholtz Society and the German Cancer Research Center (DKFZ). The CECILE study was funded by the Fondation de France, the French National Institute of Cancer (INCa), The National League against Cancer, the National Agency for Environmental and Occupational Health and Food Safety (ANSES), the National Agency for Research (ANR) and the Association for Research against Cancer (ARC). The CGPS was supported by the Chief Physician Johan Boserup and Lise

Boserup Fund, the Danish Medical Research Council and Herlev Hospital. The CNIO-BCS was supported by the Genome Spain Foundation, the Red Temática de Investigación Cooperativa en Cáncer and grants from the Asociación Española Contra el Cáncer and the Fondo de Investigación Sanitario (PI11/00923 and PI081120). The Human Genotyping-CEGEN Unit, CNIO is supported by the Instituto de Salud Carlos III. DA was supported by a Fellowship from the Michael Manzella Foundation (MMF) and was a participant in the CNIO Summer Training Program. The CTS was initially supported by the California Breast Cancer Act of 1993 and the California Breast Cancer Research Fund (contract 97-10500) and is currently funded through the National Institutes of Health (R01 CA77398). Collection of cancer incidence data was supported by the California Department of Public Health as part of the statewide cancer reporting program mandated by California Health and Safety Code Section 103885. HAC receives support from the Lon V Smith Foundation (LVS39420). The ESTHER study was supported by a grant from the Baden Württemberg Ministry of Science, Research and Arts. Additional cases were recruited in the context of the VERDI study, which was supported by a grant from the German Cancer Aid (Deutsche Krebshilfe). The GENICA was funded by the Federal Ministry of Education and Research (BMBF) Germany grants 01KW9975/5, 01KW9976/8, 01KW9977/0 and 01KW0114, the Robert Bosch Foundation, Stuttgart, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA) as well as the Department of Internal Medicine, Evangelische Kliniken Bonn GmbH, Johanniter Krankenhaus Bonn, Germany. The HEBCS was supported by the Helsinki University Central Hospital Research Fund, Academy of Finland (132473), the Finnish Cancer Society, The Nordic Cancer Union and the Sigrid Juselius Foundation. The HERPACC was supported by a Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Science, Sports, Culture and Technology of Japan, by a Grant-in-Aid for the Third-Term Comprehensive 10-Year Strategy for Cancer Control from Ministry Health, Labour and Welfare of Japan, by a research grant from Takeda Science Foundation, by Health and Labour Sciences Research Grants for Research on Applying Health Technology from Ministry Health, Labour and Welfare of Japan and by National Cancer Center Research and Development Fund. The HMBCS was supported by short-term fellowships from the German Academic Exchange Program (to NB) and the Friends of Hannover Medical School (to NB)—Financial support for KARBAC was provided through the regional agreement on medical training and clinical research (ALF) between Stockholm County Council and Karolinska Institutet, the Stockholm Cancer Foundation and the Swedish Cancer Society. The KBCP was financially supported by the special Government Funding (EVO) of Kuopio University Hospital grants, Cancer Fund of North Savo, the Finnish Cancer Organizations, the Academy of Finland and by the strategic funding of the University of Eastern Finland. kConFab is supported by grants from the National Breast Cancer Foundation, the NHMRC, the Queensland Cancer Fund, the Cancer Councils of New South Wales, Victoria, Tasmania and South Australia and the Cancer Foundation of Western Australia. The kConFab Clinical Follow-Up Study was funded by the NHMRC (145684, 288704 and 454508). Financial support for the AOCS was provided by the United States Army Medical Research and Materiel Command (DAMD17-01-1-0729), the Cancer Council of Tasmania and Cancer Foundation of Western Australia and the NHMRC (199600). GCT and PW are supported by the NHMRC. LAABC is supported by grants (1RB-0287, 3PB-0102, 5PB-0018 and 10PB-0098) from the California Breast Cancer Research Program. Incident breast cancer cases were collected by the USC Cancer Surveillance Program (CSP) which is supported under subcontract by the California Department of Health. The CSP is also part of the National Cancer Institute's Division of Cancer Prevention and Control Surveillance, Epidemiology and End Results Program, under contract number N01CN25403. LMBC is supported by the 'Stichting tegen Kanker' (232-2008 and 196-2010). The MARIE study was supported by the Deutsche Krebshilfe e.V. (70-2892-BR I), the Federal Ministry of Education and Research (BMBF) Germany (01KH0402), the Hamburg Cancer Society and the German Cancer Research Center (DKFZ). MBCSG is supported by grants from the Italian Association for Cancer Research (AIRC) and by funds from the Italian citizens who allocated a 5 out of 1000 share of their tax payment in support of the Fondazione IRCCS Istituto Nazionale dei Tumori, according to Italian laws (INT-Institutional strategic projects "5 × 1000"). The MCBCS was supported by the NIH grants (CA122340 and CA128978) and a Specialized Program of Research Excellence (SPORE) in Breast Cancer (CA116201), the Breast Cancer Research Foundation and a generous gift from the David F and Margaret T Grohne Family Foundation and the Ting Tsung and Wei Fong Chao Foundation. MCCS cohort recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS was further supported by Australian NHMRC grants 209057, 251553 and 504711 and by infrastructure provided by Cancer Council Victoria. The MEC was supported by NIH grants CA63464, CA54281, CA098758 and CA132839. The work of MTLGEBCS was supported by the Quebec Breast Cancer Foundation, the Canadian Institutes of Health Research (grant CRN-87521) and the Ministry of Economic Development, Innovation and Export Trade (grant PSR-SIIRI-701). MYBRCA is funded by research grants from the Malaysian Ministry of Science, Technology and Innovation (MOSTI), Malaysian Ministry of Higher Education (UM.C/HlR/MOHE/06) and Cancer Research Initiatives Foundation (CARIF). Additional controls were recruited by the Singapore Eye Research Institute, which was supported by a grant from the Biomedical Research Council (BMRC08/1/35/19<tel:08/1/35/19>/550), Singapore and the National medical Research Council, Singapore (NMRC/CG/SERI/2010). The NBCS was supported by grants from the Norwegian Research council (155218/V40, 175240/S10 to ALBD, FUGE-NFR 181600/V11 to VNK and a Swizz Bridge Award to ALBD). The NBHS was supported by NIH grant R01CA100374. Biological sample preparation was conducted the Survey and Biospecimen Shared Resource, which is supported by P30 CA68485. The OBCS was supported by research grants from the Finnish Cancer Foundation, the Sigrid Juselius Foundation, the Academy of Finland, the University of Oulu, and the Oulu University Hospital. The ORIGO study was supported by the Dutch Cancer Society (RUL 1997-1505) and the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL CP16). The PBCS was funded by Intramural Research Funds of the National Cancer Institute, Department of Health and Human Services, USA. pKARMA is a combination of the KARMA and LIBRO-1 studies. KARMA was supported by Märit and Hans Rausings Initiative against Breast Cancer. KARMA and LIBRO-1 were supported the Cancer Risk Prediction Center (CRisP; www.crispcenter.org), a Linnaeus Centre (Contract ID 70867902) financed by the Swedish Research Council. The RBCS was funded by the Dutch Cancer Society (DDHK 2004-3124, DDHK 2009-4318). SASBAC was supported by funding from the Agency for Science, Technology and Research of Singapore (A*STAR), the US National Institute of Health (NIH) and the Susan G Komen Breast Cancer Foundation. KC was financed by the Swedish Cancer Society (5128-B07-01PAF). The SBCGS was supported primarily by NIH grants R01CA64277, R01CA148667 and R37CA70867. Biological sample preparation was conducted the Survey and Biospecimen Shared Resource, which is supported by P30 CA68485. The SBCS was supported by Yorkshire Cancer Research S305PA, S299 and S295. Funding for the SCCS was provided by

NIH grant R01 CA092447. The Arkansas Central Cancer Registry is fully funded by a grant from National Program of Cancer Registries, Centers for Disease Control and Prevention (CDC). Data on SCCS cancer cases from Mississippi were collected by the Mississippi Cancer Registry which participates in the National Program of Cancer Registries (NPCR) of the Centers for Disease Control and Prevention (CDC). The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the CDC or the Mississippi Cancer Registry. SEARCH is funded by a programme grant from Cancer Research UK (C490/A10124) and supported by the UK National Institute for Health Research Biomedical Research Centre at the University of Cambridge. SGBCC is funded by the National Medical Research Council start-up Grant and Centre Grant (NMRC/CG/NCIS/2010). The recruitment of controls by the Singapore Consortium of Cohort Studies-Multi-ethnic cohort (SCCS-MEC) was funded by the Biomedical Research Council (grant number: 05/1/21/19/425). SKKDKFZS is supported by the DKFZ. Katarzyna Jaworska is a fellow of International PhD program, Postgraduate School of Molecular Medicine, Warsaw Medical University, supported by the Polish Foundation of Science. The TNBCC was supported by the NIH grant (CA128978), the Breast Cancer Research Foundation, Komen Foundation for the Cure, the Ohio State University Comprehensive Cancer Center, the Stefanie Spielman fund for Breast Cancer Research and a generous gift from the David F and Margaret T Grohne Family Foundation and the Ting Tsung and Wei Fong Chao Foundation. Part of the TNBCC (DEMOKRITOS) has been co-financed by the European Union (European Social Fund–ESF) and Greek national funds through the Operational Program 'Education and Lifelong Learning' of the National Strategic Reference Framework (NSRF)—Research Funding Program of the General Secretariat for Research & Technology: ARISTEIA. The TWBCS is supported by the Institute of Biomedical Sciences, Academia Sinica and the National Science Council, Taiwan. The UKBGS is funded by Breakthrough Breast Cancer and the Institute of Cancer Research (ICR). ICR acknowledges NHS funding to the NIHR Biomedical Research Centre.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

[1]Human Cancer Genetics Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain; [2]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT, USA; [3]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; [4]Biostatistics Unit, IMDEA Food Institute, Madrid, Spain; [5]Human Genotyping-CEGEN Unit, Spanish National Cancer Research Centre (CNIO), Madrid, Spain; [6]CRUK Cambridge Institute, University of Cambridge, Cambridge, UK; [7]Hospital Monte Naranco, Oviedo, Spain; [8]Department of Pathology, Hospital Universitario La Paz, IdiPAZ (Hospital La Paz Institute for Health Research) Universidad Autonoma de Madrid, Madrid, Spain; [9]Laboratory of Pathology and Oncology, Research Unit, Hospital Universitario La Paz, IdiPAZ, Madrid, Spain; [10]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden; [11]Department of Oncology—Pathology, Karolinska Institutet, Stockholm, Sweden; [12]Division of Genetics and Epidemiology, The Institute of Cancer Research, Sutton, UK; [13]Division of Breast Cancer Research, The Institute of Cancer Research, London, UK; [14]Breakthrough Breast Cancer Research Centre, Division of Breast Cancer Research, The Institute of Cancer Research, London, UK; [15]Department of Preventive Medicine, Kyushu University Faculty of Medical Sciences, Fukuoka, Japan; [16]Division of Epidemiology and Prevention, Aichi Cancer Center Research Institute, Nagoya, Japan; [17]Department of Breast Oncology, Aichi Cancer Center Hospital, Nagoya, Japan; [18]Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia; [19]QIMR Berghofer Institute of Medical Research, Brisbane, Queensland, Australia; [20]Saw Swee Hock School of Public Health, Department of Surgery, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore; [21]Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore, Singapore; [22]Department of Surgery, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore; [23]Research Oncology, Division of Cancer Studies, Kings College London Guy's Hospital, London, UK; [24]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK; [25]Oxford Biomedical Research Centre, University of Oxford, Oxford, UK; [26]School of Medicine, Clinical Science Institute, National University of Ireland, Galway, UK; [27]Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, Korea; [28]Department of Biomedical Sciences, Seoul National University Graduate School, Seoul, Korea; [29]Department of Surgery, Seoul National University College of Medicine, Seoul, Korea; [30]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Victoria, Australia; [31]Department of Pathology, University of Melbourne, Melbourne, Victoria, Australia; [32]Cancer Research Initiatives Foundation, Sime Darby Medical Centre, Subang Jaya, Malaysia; [33]Breast Cancer Research Unit, University Malaya Cancer Research Institute, University Malaya Medical Centre, Kuala Lumpur, Malaysia; [34]Singapore Eye Research Institute, National University of Singapore, Singapore; [35]Dr Margarete Fischer-Bosch Institute of Clinical Pharmacology, Stuttgart, Germany; [36]University of Tübingen, Germany; [37]Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr-University Bochum (IPA), Germany; [38]Molecular Genetics of Breast Cancer, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, Germany; [39]Institute for Occupational Medicine and Maritime Medicine, University Medical Center Hamburg-Eppendorf, Germany; [40]Institute of Pathology, Medical Faculty of the University of Bonn, Germany; [41]Department of Internal Medicine, Evangelische Kliniken Bonn GmbH, Johanniter Krankenhaus, Bonn, Germany; [42]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK; [43]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada; [44]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada; [45]Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; [46]Ontario Cancer Genetics Network, Lunenfeld-Tanenbaum Research Institute, Toronto, Ontario, Canada; [47]Netherlands Cancer Institute, Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands; [48]Department of Gynecology and Obstetrics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, Erlangen, Germany; [49]David Geffen School of Medicine, Department of Medicine Division of Hematology and Oncology, University of California at Los Angeles, CA, USA; [50]Institute of Human Genetics, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg,

Erlangen, Germany; [51]Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan; [52]College of Public Health, China Medical University, Taichong, Taiwan; [53]Tri-Service General Hospital, Taipei, Taiwan; [54]Cancer Center, Kaohsiung Medical University Chung-Ho Memorial Hospital, Kaohsiung, Taiwan; [55]Department of Surgery, Kaohsiung Medical University Chung-Ho Memorial Hospital, Kaohsiung, Taiwan; [56]Department of Medicine, Vanderbilt University, Nashville, TN USA; [57]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA; [58]CRUK/YCR Sheffield Cancer Research Centre, Department of Oncology, University of Sheffield, Sheffield, UK; [59]Institute of Population Health, University of Manchester, Manchester, UK; [60]Division of Health Sciences, Warwick Medical School, Coventry, UK; [61]Ministry of Public Health, Bangkok, Thailand; [62]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA; [63]Shanghai Center for Disease Control and Prevention, Shanghai, China; [64]Shanghai Cancer Institute, Shanghai, China; [65]Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Preventive and Predictive Medicine, Fondazione IRCCS Istituto Nazionale dei Tumori (INT), Milan, Italy; [66]IFOM, Fondazione Istituto FIRC di Oncologia Molecolare, Milan, Italy; [67]Unit of Medical Genetics, Department of Preventive and Predictive Medicine, Fondazione IRCCS Istituto Nazionale dei Tumori (INT), Milan, Italy; [68]Cogentech Cancer Genetic Test Laboratory, Milan, Italy; [69]National Cancer Institute, Bangkok, Thailand; [70]Genetic Susceptibility Group, International Agency for Research on Cancer, Lyon, France; [71]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA; [72]Department of Molecular Virology, Immunology and Medical Genetics, Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA; [73]Mayo Clinic, Rochester, MN, USA; [74]Molecular Diagnostics Laboratory, INRASTES, National Centre for Scientific Research "Demokritos", Athens, Greece; [75]Breakthrough Breast Cancer Research Centre, The Institute of Cancer Research, London, UK; [76]London School of Hygiene and Tropical Medicine, London, UK; [77]Department of Obstetrics and Gynecology, University of Heidelberg, Heidelberg, Germany; [78]National Center for Tumor Diseases, University of Heidelberg, Heidelberg, Germany; [79]Molecular Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany; [80]Inserm (National Institute of Health and Medical Research), CESP (Center for Research in Epidemiology and Population Health), U1018, Environmental Epidemiology of Cancer, Villejuif, France; [81]University Paris-Sud, UMRS 1018, Villejuif, France; [82]Inserm (National Institute of Health and Medical Research), U775 Paris, France; [83]Centre de Ressources Biologiques EPIGENETEC, Paris, France; [84]Copenhagen General Population Study, Herlev University Hospital, University of Copenhagen, Copenhagen, Denmark; [85]Department of Clinical Biochemistry, Herlev University Hospital, University of Copenhagen, Copenhagen, Denmark; [86]Department of Breast Surgery, Herlev University Hospital, University of Copenhagen, Copenhagen, Denmark; [87]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany; [88]German Cancer Consortium (DKTK), Heidelberg, Germany; [89]Saarland Cancer Registry, Saarbrücken, Germany; [90]School of Medicine, Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, Kuopio, Finland; [91]Biocenter Kuopio, University of Eastern Finland, Kuopio, Finland; [92]Department of Clinical Pathology, Kuopio University Hospital, Kuopio, Finland; [93]School of Medicine, Institute of Clinical Medicine, Oncology, University of Eastern Finland, Kuopio, Finland; [94]Cancer Center, Kuopio University Hospital, Kuopio, Finland; [95]Vesalius Research Center (VRC), VIB, Leuven, Belgium; [96]Multidisciplinary Breast Center, University Hospital Gasthuisberg, Leuven, Belgium; [97]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany; [98]Institute for Medical Biometrics and Epidemiology, University Clinic Hamburg-Eppendorf, Hamburg, Germany; [99]Department of Cancer Epidemiology/Clinical Cancer Registry, University Clinic Hamburg-Eppendorf, Hamburg, Germany; [100]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA; [101]Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, Australia; [102]University of Hawaii Cancer Center, Honolulu, HI, USA; [103]Cancer Genomics Laboratory, Centre Hospitalier Universitaire de Quebec Research Center and Laval University, Quebec, Canada; [104]Department of Medicine, McGill University, Montreal, Montreal, Quebec, Canada; [105]Division of Clinical Epidemiology, McGill University Health Centre, Royal Victoria Hospital, Montreal, Quebec, Canada; [106]Département de médecine sociale et préventive, Département de santé environnementale et santé au travail, Université de Montréal, Montreal, Quebec, Canada; [107]Laboratory of Cancer Genetics and Tumor Biology, Department of Clinical Chemistry and Biocenter Oulu, University of Oulu, Oulu University Hospital, Oulu, Finland; [108]Department of Oncology, Oulu University Hospital, University of Oulu, Oulu, Finland; [109]Department of Surgery, Oulu University Hospital, University of Oulu, Oulu, Finland; [110]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands; [111]Department of Surgical Oncology, Leiden University Medical Center, Leiden, The Netherlands; [112]Department of Medical Oncology, Family Cancer Clinic, Erasmus University Medical Centre, Rotterdam, The Netherlands; [113]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA; [114]Division of Genetics and Epidemiology, Institute of Cancer Research and Breakthrough Breast Cancer Research Centre, London, UK; [115]Department of Cancer Epidemiology and Prevention, M. Sklodowska-Curie Memorial Cancer Center & Institute of Oncology, Warsaw, Poland; [116]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [117]Department of Clinical Genetics, Family Cancer Clinic, Erasmus University Medical Center, Rotterdam, The Netherlands; [118]Department of Surgical Oncology, Family Cancer Clinic, Erasmus University Medical Centre, Rotterdam, The Netherlands; [119]Human Genetics Division, Genome Institute of Singapore, Singapore, Singapore; [120]Department of Genetics and Pathology, Pomeranian Medical University, Szczecin, Poland; [121]Department of Obstetrics and Gynecology, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland; [122]Department of Clinical Genetics, Helsinki University Central Hospital, Helsinki, Finland; [123]Department of Oncology, Helsinki University Central Hospital, Helsinki, Finland; [124]Department of Obstetrics and Gynaecology, Hannover Medical School, Hannover, Germany; [125]Department of Radiation Oncology, Hannover Medical School, Hannover, Germany; [126]Servicio de Cirugía General y Especialidades, Hospital Monte Naranco, Oviedo, Spain; [127]Servicio de Oncología Médica, Hospital Universitario La Paz, Madrid, Spain

REFERENCES

Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, Platte R, Morrison J, Maranian M, Pooley KA, Luben R, Eccles D, Evans DG, Fletcher O, Johnson N, dos Santos Silva I, Peto J, Stratton MR, Rahman N, Jacobs K, Prentice R, Anderson GL, Rajkovic A, Curb JD, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Diver WR, Bojesen S, Nordestgaard BG, Flyger H, Dork T, Schurmann P, Hillemanns P, Karstens JH, Bogdanova NV, Antonenkova NN, Zalutsky IV, Bermisheva M, Fedorova S, Khusnutdinova E, Kang D, Yoo KY, Noh DY, Ahn SH, Devilee P, van Asperen CJ, Tollenaar RA, Seynaeve C, Garcia-Closas M, Lissowska J, Brinton L, Peplonska B, Nevanlinna H, Heikkinen T, Aittomaki K, Blomqvist C, Hopper JL, Southey MC, Smith L, Spurdle AB, Schmidt MK, Broeks A, van Hien RR, Cornelissen S, Milne RL, Ribas G, Gonzalez-Neira A, Benitez J, Schmutzler RK, Burwinkel B, Bartram CR, Meindl A, Brauch H, Justenhoven C, Hamann U, Chang-Claude J, Hein R, Wang-Gohrke S, Lindblom A, Margolin S, Mannermaa A, Kosma VM, Kataja V, Olson JE, Wang X, Fredericksen Z, Giles GG, Severi G, Baglietto L, English DR, Hankinson SE, Cox DG, Kraft P, Vatten LJ, Hveem K, Kumle M, Sigurdson A, Doody M, Bhatti P, Alexander BH, Hooning MJ, van den Ouweland AM, Oldenburg RA, Schutte M, Hall P, Czene K, Liu J, Li Y, Cox A, Elliott G, Brock I, Reed MW, Shen CY, Yu JC, Hsu GC, Chen ST, Anton-Culver H, Ziogas A, Andrulis IL, Knight JA, Beesley J, Goode EL, Couch F, Chenevix-Trench G, Hoover RN, Ponder BA, Hunter DJ, Pharoah PD, Dunning AM, Chanock SJ, Easton DF (2009) Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat Genet* **41**: 585–590.

Antoniou A, Pharoah PD, Narod S, Risch HA, Eyfjord JE, Hopper JL, Loman N, Olsson H, Johannsson O, Borg A, Pasini B, Radice P, Manoukian S, Eccles DM, Tang N, Olah E, Anton-Culver H, Warner E, Lubinski J, Gronwald J, Gorski B, Tulinius H, Thorlacius S, Eerola H, Nevanlinna H, Syrjakoski K, Kallioniemi OP, Thompson D, Evans C, Peto J, Lalloo F, Evans DG, Easton DF (2003) Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. *Am J Hum Genet* **72**: 1117–1130.

Antoniou AC, Wang X, Fredericksen ZS, McGuffog L, Tarrell R, Sinilnikova OM, Healey S, Morrison J, Kartsonaki C, Lesnick T, Ghoussaini M, Barrowdale D, Peock S, Cook M, Oliver C, Frost D, Eccles D, Evans DG, Eeles R, Izatt L, Chu C, Douglas F, Paterson J, Stoppa-Lyonnet D, Houdayer C, Mazoyer S, Giraud S, Lasset C, Remenieras A, Caron O, Hardouin A, Berthet P, Hogervorst FB, Rookus MA, Jager A, van den Ouweland A, Hoogerbrugge N, van der Luijt RB, Meijers-Heijboer H, Gomez Garcia EB, Devilee P, Vreeswijk MP, Lubinski J, Jakubowska A, Gronwald J, Huzarski T, Byrski T, Gorski B, Cybulski C, Spurdle AB, Holland H, Goldgar DE, John EM, Hopper JL, Southey M, Buys SS, Daly MB, Terry MB, Schmutzler RK, Wappenschmidt B, Engel C, Meindl A, Preisler-Adams S, Arnold N, Niederacher D, Sutter C, Domchek SM, Nathanson KL, Rebbeck T, Blum JL, Piedmonte M, Rodriguez GC, Wakeley K, Boggess JF, Basil J, Blank SV, Friedman E, Kaufman B, Laitman Y, Milgrom R, Andrulis IL, Glendon G, Ozcelik H, Kirchhoff T, Vijai J, Gaudet MM, Altshuler D, Guiducci C, Loman N, Harbst K, Rantala J, Ehrencrona H, Gerdes AM, Thomassen M, Sunde L, Peterlongo P, Manoukian S, Bonanni B, Viel A, Radice P, Caldes T, de la Hoya M, Singer CF, Fink-Retter A, Greene MH, Mai PL, Loud JT, Guidugli L, Lindor NM, Hansen TV, Nielsen FC, Blanco I, Lazaro C, Garber J, Ramus SJ, Gayther SA, Phelan C, Narod S, Szabo CI, Benitez J, Osorio A, Nevanlinna H, Heikkinen T, Caligo MA, Beattie MS, Hamann U, Godwin AK, Montagna M, Casella C, Neuhausen SL, Karlan BY, Tung N, Toland AE, Weitzel J, Olopade O, Simard J, Soucy P, Rubinstein WS, Arason A, Rennert G, Martin NG, Montgomery GW, Chang-Claude J, Flesch-Janys D, Brauch H, Severi G, Baglietto L, Cox A, Cross SS, Miron P, Gerty SM, Tapper W, Yannoukakos D, Fountzilas G, Fasching PA, Beckmann MW, Dos Santos Silva I, Peto J, Lambrechts D, Paridaens R, Rudiger T, Forsti A, Winqvist R, Pylkas K, Diasio RB, Lee AM, Eckel-Passow J, Vachon C, Blows F, Driver K, Dunning A, Pharoah PP, Offit K, Pankratz VS, Hakonarson H, Chenevix-Trench G, Easton DF, Couch FJ (2010) A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. *Nat Genet* **42**: 885–892.

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**: 1294–1296.

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**: 263–265.

Bojesen SE, Pooley KA, Johnatty SE, Beesley J, Michailidou K, Tyrer JP, Edwards SL, Pickett HA, Shen HC, Smart CE, Hillman KM, Mai PL, Lawrenson K, Stutz MD, Lu Y, Karevan R, Woods N, Johnston RL, French JD, Chen X, Weischer M, Nielsen SF, Maranian MJ, Ghoussaini M, Ahmed S, Baynes C, Bolla MK, Wang Q, Dennis J, McGuffog L, Barrowdale D, Lee A, Healey S, Lush M, Tessier DC, Vincent D, Bacot F, Vergote I, Lambrechts S, Despierre E, Risch HA, Gonzalez-Neira A, Rossing MA, Pita G, Doherty JA, Alvarez N, Larson MC, Fridley BL, Schoof N, Chang-Claude J, Cicek MS, Peto J, Kalli KR, Broeks A, Armasu SM, Schmidt MK, Braaf LM, Winterhoff B, Nevanlinna H, Konecny GE, Lambrechts D, Rogmann L, Guenel P, Teoman A, Milne RL, Garcia JJ, Cox A, Shridhar V, Burwinkel B, Marme F, Hein R, Sawyer EJ, Haiman CA, Wang-Gohrke S, Andrulis IL, Moysich KB, Hopper JL, Odunsi K, Lindblom A, Giles GG, Brenner H, Simard J, Lurie G, Fasching PA, Carney ME, Radice P, Wilkens LR, Swerdlow A, Goodman MT, Brauch H, Garcia-Closas M, Hillemanns P, Winqvist R, Durst M, Devilee P, Runnebaum I, Jakubowska A, Lubinski J, Mannermaa A, Butzow R, Bogdanova NV, Dork T, Pelttari LM, Zheng W, Leminen A, Anton-Culver H, Bunker CH, Kristensen V, Ness RB, Muir K, Edwards R, Meindl A, Heitz F, Matsuo K, du Bois A, Wu AH, Harter P, Teo SH, Schwaab I, Shu XO, Blot W, Hosono S, Kang D, Nakanishi T, Hartman M, Yatabe Y, Hamann U, Karlan BY, Sangrajrang S, Kjaer SK, Gaborieau V, Jensen A, Eccles D, Hogdall E, Shen CY, Brown J, Woo YL, Shah M, Azmi MA, Luben R, Omar SZ, Czene K, Vierkant RA, Nordestgaard BG, Flyger H, Vachon C, Olson JE, Wang X, Levine DA, Rudolph A, Weber RP, Flesch-Janys D, Iversen E, Nickels S, Schildkraut JM, Silva Idos S, Cramer DW, Gibson L, Terry KL, Fletcher O, Vitonis AF, van der Schoot CE, Poole EM, Hogervorst FB, Tworoger SS, Liu J, Bandera EV, Li J, Olson SH, Humphreys K, Orlow I, Blomqvist C, Rodriguez-Rodriguez L, Aittomaki K, Salvesen HB, Muranen TA, Wik E, Brouwers B, Krakstad C, Wauters E, Halle MK, Wildiers H, Kiemeney LA, Mulot C, Aben KK, Laurent-Puig P, Altena AM, Truong T, Massuger LF, Benitez J, Pejovic T, Perez JI, Hoatlin M, Zamora MP, Cook LS, Balasubramanian SP, Kelemen LE, Schneeweiss A, Le ND, Sohn C, Brooks-Wilson A, Tomlinson I, Kerin MJ, Miller N, Cybulski C, Henderson BE, Menkiszak J, Schumacher F, Wentzensen N, Le Marchand L, Yang HP, Mulligan AM, Glendon G, Engelholm SA, Knight JA, Hogdall CK, Apicella C, Gore M, Tsimiklis H, Song H, Southey MC, Jager A, den Ouweland AM, Brown R, Martens JW, Flanagan JM, Kriege M, Paul J, Margolin S, Siddiqui N, Severi G, Whittemore AS, Baglietto L, McGuire V, Stegmaier C, Sieh W, Muller H, Arndt V, Labreche F, Gao YT, Goldberg MS, Yang G, Dumont M, McLaughlin JR, Hartmann A, Ekici AB, Beckmann MW, Phelan CM, Lux MP, Permuth-Wey J, Peissel B, Sellers TA, Ficarazzi F, Barile M, Ziogas A, Ashworth A, Gentry-Maharaj A, Jones M, Ramus SJ, Orr N, Menon U, Pearce CL, Bruning T, Pike MC, Ko YD, Lissowska J, Figueroa J, Kupryjanczyk J, Chanock SJ, Dansonka-Mieszkowska A, Jukkola-Vuorinen A, Rzepecka IK, Pylkas K, Bidzinski M, Kauppila S, Hollestelle A, Seynaeve C, Tollenaar RA, Durda K, Jaworska K, Hartikainen JM, Kosma VM, Kataja V, Antonenkova NN, Long J, Shrubsole M, Deming-Halverson S, Lophatananon A, Siriwanarangsan P, Stewart-Brown S, Ditsch N, Lichtner P, Schmutzler RK, Ito H, Iwata H, Tajima K, Tseng CC, Stram DO, van den Berg D, Yip CH, Ikram MK, Teh YC, Cai H, Lu W, Signorello LB, Cai Q, Noh DY, Yoo KY, Miao H, Iau PT, Teo YY, McKay J, Shapiro C, Ademuyiwa F, Fountzilas G, Hsiung CN, Yu JC, Hou MF, Healey CS, Luccarini C, Peock S, Stoppa-Lyonnet D, Peterlongo P, Rebbeck TR, Piedmonte M, Singer CF, Friedman E, Thomassen M, Offit K, Hansen TV, Neuhausen SL, Szabo CI, Blanco I, Garber J, Narod SA, Weitzel JN, Montagna M, Olah E, Godwin AK, Yannoukakos D, Goldgar DE, Caldes T, Imyanitov EN, Tihomirova L, Arun BK, Campbell I, Mensenkamp AR, van Asperen CJ, van Roozendaal KE, Meijers-Heijboer H, Collee JM, Oosterwijk JC, Hooning MJ, Rookus MA, van der Luijt RB, Os TA, Evans DG, Frost D, Fineberg E, Barwell J, Walker L, Kennedy MJ, Platte R, Davidson R, Ellis SD, Cole T, Bressac-de Paillerets B, Buecher B, Damiola F, Faivre L, Frenay M, Sinilnikova OM, Caron O, Giraud S, Mazoyer S, Bonadona V, Caux-Moncoutier V, Toloczko-Grabarek A, Gronwald J, Byrski T, Spurdle AB, Bonanni B, Zaffaroni D, Giannini G, Bernard L, Dolcetti R, Manoukian S, Arnold N, Engel C, Deissler H, Rhiem K, Niederacher D, Plendl H, Sutter C, Wappenschmidt B, Borg A, Melin B, Rantala J, Soller M, Nathanson KL, Domchek SM, Rodriguez GC, Salani R, Kaulich DG, Tea MK, Paluch SS, Laitman Y, Skytte AB, Kruse TA, Jensen UB, Robson M, Gerdes AM, Ejlertsen B, Foretova L, Savage SA, Lester J, Soucy P, Kuchenbaecker KB, Olswold C, Cunningham JM, Slager S, Pankratz VS, Dicks E, Lakhani SR, Couch FJ, Hall P, Monteiro AN, Gayther SA, Pharoah PD, Reddel RR, Goode EL, Greene MH, Easton DF, Berchuck A, Antoniou AC, Chenevix-Trench G, Dunning AM (2013) Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nat Genet* **45:** 371–384, 384. e1–e2.

Cai Q, Long J, Lu W, Qu S, Wen W, Kang D, Lee JY, Chen K, Shen H, Shen CY, Sung H, Matsuo K, Haiman CA, Khoo US, Ren Z, Iwasaki M, Gu K, Xiang YB, Choi JY, Park SK, Zhang L, Hu Z, Wu PE, Noh DY, Tajima K, Henderson BE, Chan KY, Su F, Kasuga Y, Wang W, Cheng JR, Yoo KY, Zheng H, Liu Y, Shieh YL, Kim SW, Lee JW, Iwata H, Le Marchand L, Chan SY, Xie X, Tsugane S, Lee MH, Wang S, Li G, Levy S, Huang B, Shi J, Delahanty R, Zheng Y, Li C, Gao YT, Shu XO, Zheng W (2011) Genome-wide association study identifies breast cancer risk variant at 10q21.2: results from the Asia Breast Cancer Consortium. *Hum Mol Genet* **20**: 4991–4999.

Chen GJ, Forough R (2006) Fibroblast growth factors, fibroblast growth factor receptors, diseases, and drugs. *Recent Pat Cardiovasc Drug Discov* **1**: 211–224.

Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, Wareham N, Ahmed S, Healey CS, Bowman R, Meyer KB, Haiman CA, Kolonel LK, Henderson BE, Le Marchand L, Brennan P, Sangrajrang S, Gaborieau V, Odefrey F, Shen CY, Wu PE, Wang HC, Eccles D, Evans DG, Peto J, Fletcher O, Johnson N, Seal S, Stratton MR, Rahman N, Chenevix-Trench G, Bojesen SE, Nordestgaard BG, Axelsson CK, Garcia-Closas M, Brinton L, Chanock S, Lissowska J, Peplonska B, Nevanlinna H, Fagerholm R, Eerola H, Kang D, Yoo KY, Noh DY, Ahn SH, Hunter DJ, Hankinson SE, Cox DG, Hall P, Wedren S, Liu J, Low YL, Bogdanova N, Schurmann P, Dork T, Tollenaar RA, Jacobi CE, Devilee P, Klijn JG, Sigurdson AJ, Doody MM, Alexander BH, Zhang J, Cox A, Brock IW, MacPherson G, Reed MW, Couch FJ, Goode EL, Olson JE, Meijers-Heijboer H, van den Ouweland A, Uitterlinden A, Rivadeneira F, Milne RL, Ribas G, Gonzalez-Neira A, Benitez J, Hopper JL, McCredie M, Southey M, Giles GG, Schroen C, Justenhoven C, Brauch H, Hamann U, Ko YD, Spurdle AB, Beesley J, Chen X, Mannermaa A, Kosma VM, Kataja V, Hartikainen J, Day NE, Cox DR, Ponder BA (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**: 1087–1093.

Eswarakumar VP, Lax I, Schlessinger J (2005) Cellular signaling by fibroblast growth factor receptors. *Cytokine Growth Factor Rev* **16**: 139–149.

Fletcher O, Johnson N, Orr N, Hosking FJ, Gibson LJ, Walker K, Zelenika D, Gut I, Heath S, Palles C, Coupland B, Broderick P, Schoemaker M, Jones M, Williamson J, Chilcott-Burns S, Tomczyk K, Simpson G, Jacobs KB, Chanock SJ, Hunter DJ, Tomlinson IP, Swerdlow A, Ashworth A, Ross G, dos Santos Silva I, Lathrop M, Houlston RS, Peto J (2011) Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. *J Natl Cancer Inst* **103**: 425–435.

Garcia-Closas M, Couch FJ, Lindstrom S, Michailidou K, Schmidt MK, Brook MN, Orr N, Rhie SK, Riboli E, Feigelson HS, Le Marchand L, Buring JE, Eccles D, Miron P, Fasching PA, Brauch H, Chang-Claude J, Carpenter J, Godwin AK, Nevanlinna H, Giles GG, Cox A, Hopper JL, Bolla MK, Wang Q, Dennis J, Dicks E, Howat WJ, Schoof N, Bojesen SE, Lambrechts D, Broeks A, Andrulis IL, Guenel P, Burwinkel B, Sawyer EJ, Hollestelle A, Fletcher O, Winqvist R, Brenner H, Mannermaa A, Hamann U, Meindl A, Lindblom A, Zheng W, Devillee P, Goldberg MS, Lubinski J, Kristensen V, Swerdlow A, Anton-Culver H, Dork T, Muir K, Matsuo K, Wu AH, Radice P, Teo SH, Shu XO, Blot W, Kang D, Hartman M, Sangrajrang S, Shen CY, Southey MC, Park DJ, Hammet F, Stone J, Veer LJ, Rutgers EJ, Lophatananon A, Stewart-Brown S, Siriwanarangsan P, Peto J, Schrauder MG, Ekici AB, Beckmann MW, Dos Santos Silva I, Johnson N, Warren H, Tomlinson I, Kerin MJ, Miller N, Marme F, Schneeweiss A, Sohn C, Truong T, Laurent-Puig P, Kerbrat P, Nordestgaard BG, Nielsen SF, Flyger H, Milne RL, Perez JI, Menendez P, Muller H, Arndt V, Stegmaier C, Lichtner P, Lochmann M, Justenhoven C, Ko YD, Muranen TA, Aittomaki K, Blomqvist C, Greco D, Heikkinen T, Ito H, Iwata H, Yatabe Y, Antonenkova NN, Margolin S, Kataja V, Kosma VM, Hartikainen JM, Balleine R, Tseng CC, Berg DV, Stram DO, Neven P, Dieudonne AS, Leunen K, Rudolph A, Nickels S, Flesch-Janys D, Peterlongo P, Peissel B, Bernard L, Olson JE, Wang X, Stevens K, Severi G, Baglietto L, McLean C, Coetzee GA, Feng Y, Henderson BE, Schumacher F, Bogdanova NV, Labreche F, Dumont M, Yip CH, Taib NA, Cheng CY, Shrubsole M, Long J, Pylkas K, Jukkola-Vuorinen A, Kauppila S, Knight JA, Glendon G, Mulligan AM, Tollenaar RA, Seynaeve CM, Kriege M, Hooning MJ, van den Ouweland AM, van Deurzen CH, Lu W, Gao YT, Cai H, Balasubramanian SP, Cross SS, Reed MW, Signorello L, Cai Q, Shah M, Miao H, Chan CW, Chia KS, Jakubowska A, Jaworska K, Durda K, Hsiung CN, Wu PE, Yu JC, Ashworth A, Jones M, Tessier DC, Gonzalez-Neira A, Pita G, Alonso MR, Vincent D, Bacot F, Ambrosone CB, Bandera EV, John EM, Chen GK, Hu JJ, Rodriguez-Gil JL, Bernstein L, Press MF, Ziegler RG, Millikan RM, Deming-Halverson SL, Nyante S, Ingles SA, Waisfisz Q, Tsimiklis H, Makalic E, Schmidt D, Bui M, Gibson L, Muller-Myhsok B, Schmutzler RK, Hein R, Dahmen N, Beckmann L, Aaltonen K, Czene K, Irwanto A, Liu J, Turnbull C, Rahman N, Meijers-Heijboer H, Uitterlinden AG, Rivadeneira F, Olswold C, Slager S, Pilarski R, Ademuyiwa F, Konstantopoulou I, Martin NG, Montgomery GW, Slamon DJ, Rauh C, Lux MP, Jud SM, Bruning T, Weaver J, Sharma P, Pathak H, Tapper W, Gerty S, Durcan L, Trichopoulos D, Tumino R, Peeters PH, Kaaks R, Campa D, Canzian F, Weiderpass E, Johansson M, Khaw KT, Travis R, Clavel-Chapelon F, Kolonel LN, Chen C, Beck A, Hankinson SE, Berg CD, Hoover RN, Lissowska J, Figueroa JD, Chasman DI, Gaudet MM, Diver WR, Willett WC, Hunter DJ, Simard J, Benitez J, Dunning AM, Sherman ME, Chenevix-Trench G, Chanock SJ, Hall P, Pharoah PD, Vachon C, Easton DF, Haiman CA, Kraft P (2013) Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nat Genet* **45:** 392–398, 398. e1–e2.

Garcia-Closas M, Hall P, Nevanlinna H, Pooley K, Morrison J, Richesson DA, Bojesen SE, Nordestgaard BG, Axelsson CK, Arias JI, Milne RL, Ribas G, Gonzalez-Neira A, Benitez J, Zamora P, Brauch H, Justenhoven C, Hamann U, Ko YD, Bruening T, Haas S, Dork T, Schurmann P, Hillemanns P, Bogdanova N, Bremer M, Karstens JH, Fagerholm R, Aaltonen K, Aittomaki K, von Smitten K, Blomqvist C, Mannermaa A, Uusitupa M, Eskelinen M, Tengstrom M, Kosma VM, Kataja V, Chenevix-Trench G, Spurdle AB, Beesley J, Chen X, Devilee P, van Asperen CJ, Jacobi CE, Tollenaar RA, Huijts PE, Klijn JG, Chang-Claude J, Kropp S, Slanger T, Flesch-Janys D, Mutschelknauss E, Salazar R, Wang-Gohrke S, Couch F, Goode EL, Olson JE, Vachon C, Fredericksen ZS, Giles GG, Baglietto L, Severi G, Hopper JL, English DR, Southey MC, Haiman CA, Henderson BE, Kolonel LN, Le Marchand L, Stram DO, Hunter DJ, Hankinson SE, Cox DG, Tamimi R, Kraft P, Sherman ME, Chanock SJ, Lissowska J, Brinton LA, Peplonska B, Hooning MJ, Meijers-Heijboer H, Collee JM, van den Ouweland A, Uitterlinden AG, Liu J, Lin LY, Yuqing L, Humphreys K, Czene K, Cox A, Balasubramanian SP, Cross SS, Reed MW, Blows F, Driver K, Dunning A, Tyrer J, Ponder BA, Sangrajrang S, Brennan P, McKay J, Odefrey F, Gabrieau V, Sigurdson A, Doody M, Struewing JP, Alexander B, Easton DF, Pharoah PD (2008) Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. *PLoS Genet* **4**: e1000054.

Ghoussaini M, Fletcher O, Michailidou K, Turnbull C, Schmidt MK, Dicks E, Dennis J, Wang Q, Humphreys MK, Luccarini C, Baynes C, Conroy D, Maranian M, Ahmed S, Driver K, Johnson N, Orr N, dos Santos Silva I, Waisfisz Q, Meijers-Heijboer H, Uitterlinden AG, Rivadeneira F, Hall P, Czene K, Irwanto A, Liu J, Nevanlinna H, Aittomaki K, Blomqvist C, Meindl A, Schmutzler RK, Muller-Myhsok B, Lichtner P, Chang-Claude J, Hein R, Nickels S, Flesch-Janys D, Tsimiklis H, Makalic E, Schmidt D, Bui M, Hopper JL, Apicella C, Park DJ, Southey M, Hunter DJ, Chanock SJ, Broeks A, Verhoef S, Hogervorst FB, Fasching PA, Lux MP, Beckmann MW, Ekici AB, Sawyer E, Tomlinson I, Kerin M, Marme F, Schneeweiss A, Sohn C, Burwinkel B, Guenel P, Truong T, Cordina-Duverger E, Menegaux F, Bojesen SE, Nordestgaard BG, Nielsen SF, Flyger H, Milne RL, Alonso MR, Gonzalez-Neira A, Benitez J, Anton-Culver H, Ziogas A, Bernstein L, Dur CC, Brenner H, Muller H, Arndt V, Stegmaier C, Justenhoven C, Brauch H, Bruning T, Wang-Gohrke S, Eilber U, Dork T, Schurmann P, Bremer M, Hillemanns P, Bogdanova NV, Antonenkova NN, Rogov YI, Karstens JH, Bermisheva M, Prokofieva D, Khusnutdinova E, Lindblom A, Margolin S, Mannermaa A, Kataja V, Kosma VM, Hartikainen JM, Lambrechts D, Yesilyurt BT, Floris G, Leunen K, Manoukian S, Bonanni B, Fortuzzi S, Peterlongo P, Couch FJ, Wang X, Stevens K, Lee A, Giles GG, Baglietto L, Severi G, McLean C, Alnaes GG, Kristensen V, Borrensen-Dale AL, John EM, Miron A, Winqvist R, Pylkas K, Jukkola-Vuorinen A, Kauppila S, Andrulis IL, Glendon G, Mulligan AM, Devilee P, van Asperen CJ, Tollenaar RA, Seynaeve C, Figueroa JD, Garcia-Closas M, Brinton L, Lissowska J, Hooning MJ, Hollestelle A, Oldenburg RA, van den Ouweland AM, Cox A, Reed MW, Shah M, Jakubowska A, Lubinski J, Jaworska K, Durda K, Jones M, Schoemaker M, Ashworth A, Swerdlow A, Beesley J, Chen X, Muir KR, Lophatananon A, Rattanamongkongul S, Chaiwerawattana A, Kang D, Yoo KY, Noh DY, Shen CY, Yu JC, Wu PE, Hsiung CN, Perkins A, Swann R, Velentzis L, Eccles DM, Tapper WJ, Gerty SM, Graham NJ, Ponder BA, Chenevix-Trench G, Pharoah PD, Lathrop M, Dunning AM, Rahman N, Peto J, Easton DF (2012) Genome-wide association analysis identifies three new breast cancer susceptibility loci. *Nat Genet* **44**: 312–318.

Haiman CA, Chen GK, Vachon CM, Canzian F, Dunning A, Millikan RC, Wang X, Ademuyiwa F, Ahmed S, Ambrosone CB, Baglietto L, Balleine R, Bandera EV, Beckmann MW, Berg CD, Bernstein L, Blomqvist C, Blot WJ, Brauch H, Buring JE, Carey LA, Carpenter JE, Chang-Claude J, Chanock SJ, Chasman DI, Clarke CL, Cox A, Cross SS, Deming SL, Diasio RB, Dimopoulos AM, Driver WR, Dunnebier T, Durcan L, Eccles D, Edlund CK, Ekici AB, Fasching PA, Feigelson HS, Flesch-Janys D, Fostira F, Forsti A, Fountzilas G, Gerty SM, Giles GG, Godwin AK, Goodfellow P, Graham N, Greco D, Hamann U, Hankinson SE, Hartmann A, Hein R, Heinz J, Holbrook A, Hoover RN, Hu JJ, Hunter DJ, Ingles SA, Irwanto A, Ivanovich J, John EM, Johnson N, Jukkola-Vuorinen A, Kaaks R, Ko YD, Kolonel LN, Konstantopoulou I, Kosma VM, Kulkarni S, Lambrechts D, Lee AM, Marchand LL, Lesnick T, Liu J, Lindstrom S, Mannermaa A, Margolin S, Martin NG, Miron P, Montgomery GW, Nevanlinna H, Nickels S, Nyante S, Olswold C, Palmer J, Pathak H, Pectasides D, Perou CM, Peto J, Pharoah PD, Pooler LC, Press MF, Pylkas K, Rebbeck TR, Rodriguez-Gil JL, Rosenberg L, Ross E, Rudiger T, Silva Idos S, Sawyer E, Schmidt MK, Schulz-Wendtland R, Schumacher F, Severi G, Sheng X, Signorello LB, Sinn HP, Stevens KN, Southey MC, Tapper WJ, Tomlinson I, Hogervorst FB, Wauters E, Weaver J, Wildiers H, Winqvist R, Van Den Berg D, Wan P, Xia LY, Yannoukakos D, Zheng W, Ziegler RG, Siddiq A, Slager SL, Stram DO, Easton D, Kraft P, Henderson BE, Couch FJ (2011) A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. *Nat Genet* **43**: 1210–1214.

Haldane JBS (1954) An exact test for randomness of mating. *J Genet* **52**: 631–635.

Higgins JP, Thompson SG (2002) Quantifying heterogeneity in a meta-analysis. *Stat Med* **21**: 1539–1558.

Huijts PE, van Dongen M, de Goeij MC, van Moolenbroek AJ, Blanken F, Vreeswijk MP, de Kruijf EM, Mesker WE, van Zwet EW, Tollenaar RA, Smit VT, van Asperen CJ, Devilee P (2011) Allele-specific regulation of FGFR2 expression is cell type-dependent and may increase breast cancer risk through a paracrine stimulus involving FGF10. *Breast Cancer Res* **13**: R72.

Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Hayes RB, Tucker M, Gerhard DS, Fraumeni Jr JF, Hoover RN, Thomas G, Chanock SJ (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* **39**: 870–874.

Li J, Ji L (2005) Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)* **95**: 221–227.

Meijers-Heijboer H, van den Ouweland A, Klijn J, Wasielewski M, de Snoo A, Oldenburg R, Hollestelle A, Houben M, Crepin E, van Veghel-Plandsoen M, Elstrodt F, van Duijn C, Bartels C, Meijers C, Schutte M, McGuffog L, Thompson D, Easton D, Sodha N, Seal S, Barfoot R, Mangion J, Chang-Claude J, Eccles D, Eeles R, Evans DG, Houlston R, Murday V, Narod S, Peretz T, Peto J, Phelan C, Zhang HX, Szabo C, Devilee P, Goldgar D, Futreal PA, Nathanson KL, Weber B, Rahman N, Stratton MR (2002) Low-penetrance susceptibility to breast cancer due to CHEK2(*) 1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat Genet* **31**: 55–59.

Meyer KB, Maia AT, O'Reilly M, Teschendorff AE, Chin SF, Caldas C, Ponder BA (2008) Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer. *PLoS Biol* **6**: e108.

Meyer KB, O'Reilly M, Michailidou K, Carlebur S, Edwards SL, French JD, Prathalingham R, Dennis J, Bolla MK, Wang Q, de Santiago I, Hopper JL, Tsimiklis H, Apicella C, Southey MC, Schmidt MK, Broeks A, Van 't Veer LJ, Hogervorst FB, Muir K, Lophatananon A, Stewart-Brown S, Siriwanarangsan P, Fasching PA, Lux MP, Ekici AB, Beckmann MW, Peto J, Dos Santos Silva I, Fletcher O, Johnson N, Sawyer EJ, Tomlinson I, Kerin MJ, Miller N, Marme F, Schneeweiss A, Sohn C, Burwinkel B, Guénel P, Truong T, Laurent-Puig P, Menegaux F, Bojesen SE, Nordestgaard BG, Nielsen SF, Flyger H, Milne RL, Zamora MP, Arias JI, Benitez J, Neuhausen S, Anton-Culver H, Ziogas A, Dur CC, Brenner H, Müller H, Arndt V, Stegmaier C, Meindl A, Schmutzler RK, Engel C, Ditsch N, Brauch H, Brüning T, Ko YD, The GENICA Network, Nevanlinna H, Muranen TA, Aittomäki K, Blomqvist C, Matsuo K, Ito H, Iwata H, Yatabe Y, Dörk T, Helbig S, Bogdanova NV, Lindblom A, Margolin S, Mannermaa A, Kataja V, Kosma VM, Hartikainen JM, Chenevix-Trench G, kConFab Investigators, Australian Ovarian Cancer Study Group, Wu AH, Tseng CC, Van Den Berg D, Stram DO, Lambrechts D, Thienpont B, Christiaens MR, Smeets A, Chang-Claude J, Rudolph A, Seibold P, Flesch-Janys D, Radice P, Peterlongo P, Bonanni B, Bernard L, Couch FJ, Olson JE, Wang X, Purrington K, Giles GG, Severi G, Baglietto L, McLean C, Haiman CA, Henderson BE, Schumacher F, Le Marchand L, Simard J, Goldberg MS, Labrèche F, Dumont M, Teo SH, Yip CH, Phuah SY, Kristensen V, Grenaker Alnæs G, Børresen-Dale AL, Zheng W, Deming-Halverson S, Shrubsole M, Long J, Winqvist R, Pylkäs K, Jukkola-Vuorinen A, Kauppila S, Andrulis IL, Knight JA, Glendon G, Tchatchou S, Devilee P, Tollenaar RA, Seynaeve CM, García-Closas M, Figueroa J, Chanock SJ, Lissowska J, Czene K, Darabi H, Eriksson K, Hooning MJ, Martens JW, van den Ouweland AM, van Deurzen CH, Hall P, Li J, Liu J, Humphreys K, Shu XO, Lu W, Gao YT, Cai H, Cox A, Reed MW, Blot W, Signorello LB, Cai Q, Pharoah PD, Ghoussaini M, Harrington P, Tyrer J, Kang D, Choi JY, Park SK, Noh DY, Hartman M, Hui M, Lim WY, Buhari SA, Hamann U, Försti A, Rüdiger T, Ulmer HU, Jakubowska A, Lubinski J, Jaworska K, Durda K, Sangrajrang S, Gaborieau V, Brennan P, McKay J, Vachon C, Slager S, Fostira F, Pilarski R, Shen CY, Hsiung CN, Wu PE, Hou MF, Swerdlow A, Ashworth A, Orr N, Schoemaker MJ, Ponder BA, Dunning AM, Easton DF (2013) Fine-scale mapping of the FGFR2 breast cancer risk locus: putative functional variants differentially bind FOXA1 and E2F1. *Am J Hum Genet* **93**(6): 1046–1060.

Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, Schmidt MK, Chang-Claude J, Bojesen SE, Bolla MK, Wang Q, Dicks E, Lee A, Turnbull C, Rahman N, Fletcher O, Peto J, Gibson L, Dos Santos Silva I, Nevanlinna H, Muranen TA, Aittomaki K, Blomqvist C, Czene K, Irwanto A, Liu J, Waisfisz Q, Meijers-Heijboer H, Adank M, van der Luijt RB, Hein R, Dahmen N, Beckman L, Meindl A, Schmutzler RK, Muller-Myhsok B, Lichtner P, Hopper JL, Southey MC, Makalic E, Schmidt DF, Uitterlinden AG, Hofman A, Hunter DJ, Chanock SJ, Vincent D, Bacot F, Tessier DC, Canisius S, Wessels LF, Haiman CA, Shah M, Luben R, Brown J, Luccarini C, Schoof N, Humphreys K, Li J, Nordestgaard BG, Nielsen SF, Flyger H, Couch FJ, Wang X, Vachon C, Stevens KN, Lambrechts D, Moisse M, Paridaens R, Christiaens MR, Rudolph A, Nickels S, Flesch-Janys D, Johnson N, Aitken Z, Aaltonen K, Heikkinen T, Broeks A, Veer LJ, van der Schoot CE, Guenel P, Truong T, Laurent-Puig P, Menegaux F, Marme F, Schneeweiss A, Sohn C, Burwinkel B, Zamora MP, Perez JI, Pita G, Alonso MR, Cox A, Brock IW, Cross SS, Reed MW, Sawyer EJ, Tomlinson I, Kerin MJ, Miller N, Henderson BE, Schumacher F, Le Marchand L, Andrulis IL, Knight JA, Glendon G, Mulligan AM, Lindblom A, Margolin S, Hooning MJ, Hollestelle A, van den Ouweland AM, Jager A, Bui QM, Stone J, Dite GS, Apicella C, Tsimiklis H, Giles GG, Severi G, Baglietto L, Fasching PA, Haeberle L, Ekici AB, Beckmann MW, Brenner H, Muller H, Arndt V, Stegmaier C, Swerdlow A, Ashworth A, Orr N, Jones M, Figueroa J, Lissowska J, Brinton L, Goldberg MS, Labreche F, Dumont M, Winqvist R, Pylkas K, Jukkola-Vuorinen A, Grip M, Brauch H, Hamann U, Bruning T, Radice P, Peterlongo P, Manoukian S, Bonanni B, Devilee P, Tollenaar RA, Seynaeve C, van Asperen CJ, Jakubowska A, Lubinski J, Jaworska K, Durda K, Mannermaa A, Kataja V, Kosma VM, Hartikainen JM, Bogdanova NV, Antonenkova NN, Dork T, Kristensen VN, Anton-Culver H, Slager S, Toland AE, Edge S, Fostira F, Kang D, Yoo KY, Noh DY, Matsuo K, Ito H, Iwata H, Sueta A, Wu AH, Tseng CC, Van Den Berg D, Stram DO, Shu XO, Lu W, Gao YT, Cai H, Teo SH, Yip CH, Phuah SY, Cornes BK, Hartman M, Miao H, Lim WY, Sng JH, Muir K, Lophatananon A, Stewart-Brown S, Siriwanarangsan P, Shen CY, Hsiung CN, Wu PE, Ding SL, Sangrajrang S, Gaborieau V, Brennan P, McKay J, Blot WJ, Signorello LB, Cai Q, Zheng W, Deming-Halverson S, Shrubsole M, Long J, Simard J, Garcia-Closas M, Pharoah PD, Chenevix-Trench G, Dunning AM, Benitez J, Easton DF (2013) Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* **45:** 353–361, 361. e1–e2.

Nyholt DR (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* **74**: 765–769.

Presta M, Dell'Era P, Mitola S, Moroni E, Ronca R, Rusnati M (2005) Fibroblast growth factor/fibroblast growth factor receptor system in angiogenesis. *Cytokine Growth Factor Rev* **16**: 159–178.

Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, Reid S, Spanova K, Barfoot R, Chagtai T, Jayatilake H, McGuffog L, Hanks S, Evans DG, Eccles D, Easton DF, Stratton MR (2007) PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat Genet* **39**: 165–167.

Robertson A, Hill WG (1984) Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. *Genetics* **107**: 703–718.

Schwertfeger KL (2009) Fibroblast growth factors in development and cancer: insights from the mammary and prostate glands. *Curr Drug Targets* **10**: 632–644.

Seal S, Thompson D, Renwick A, Elliott A, Kelly P, Barfoot R, Chagtai T, Jayatilake H, Ahmed M, Spanova K, North B, McGuffog L, Evans DG, Eccles D, Easton DF, Stratton MR, Rahman N (2006) Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nat Genet* **38**: 1239–1241.

Siddiq A, Couch FJ, Chen GK, Lindstrom S, Eccles D, Millikan RC, Michailidou K, Stram DO, Beckmann L, Rhie SK, Ambrosone CB, Aittomaki K, Amiano P, Apicella C, Baglietto L, Bandera EV, Beckmann MW, Berg CD, Bernstein L, Blomqvist C, Brauch H, Brinton L, Bui QM, Buring JE, Buys SS, Campa D, Carpenter JE, Chasman DI, Chang-Claude J, Chen C, Clavel-Chapelon F, Cox A, Cross SS, Czene K, Deming SL, Diasio RB, Diver WR, Dunning AM, Durcan L, Ekici AB, Fasching PA, Feigelson HS, Fejerman L, Figueroa JD, Fletcher O, Flesch-Janys D, Gaudet MM, Gerty SM, Rodriguez-Gil JL, Giles GG, van Gils CH, Godwin AK, Graham N, Greco D, Hall P, Hankinson SE, Hartmann A, Hein R, Heinz J, Hoover RN, Hopper JL, Hu JJ, Huntsman S, Ingles SA, Irwanto A, Isaacs C, Jacobs KB, John EM, Justenhoven C, Kaaks R, Kolonel LN, Coetzee GA, Lathrop M, Le Marchand L, Lee AM, Lee IM, Lesnick T, Lichtner P, Liu J, Lund E, Makalic E, Martin NG, McLean CA, Meijers-Heijboer H, Meindl A, Miron P, Monroe KR, Montgomery GW, Muller-Myhsok B, Nickels S, Nyante SJ, Olswold C, Overvad K, Palli D, Park DJ, Palmer JR, Pathak H, Peto J, Pharoah P, Rahman N, Rivadeneira F, Schmidt DF, Schmutzler RK, Slager S, Southey MC, Stevens KN, Sinn HP, Press MF, Ross E, Riboli E, Ridker PM, Schumacher FR, Severi G, Dos Santos Silva I, Stone J, Sund M, Tapper WJ, Thun MJ, Travis RC, Turnbull C, Uitterlinden AG, Waisfisz Q, Wang X, Wang Z, Weaver J, Schulz-Wendtland R, Wilkens LR, Van Den Berg D, Zheng W, Ziegler RG, Ziv E, Nevanlinna H, Easton DF, Hunter DJ, Henderson BE, Chanock SJ, Garcia-Closas M, Kraft P, Haiman CA, Vachon CM (2012) A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11. *Hum Mol Genet* **21**: 5373–5384.

Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, Masson G, Jakobsdottir M, Thorlacius S, Helgason A, Aben KK, Strobbe LJ, Albers-Akkers MT, Swinkels DW, Henderson BE, Kolonel LN, Le Marchand L, Millastre E, Andres R, Godino J, Garcia-Prats MD, Polo E, Tres A, Mouy M, Saemundsdottir J, Backman VM, Gudmundsson L, Kristjansson K, Bergthorsson JT, Kostic J, Frigge ML, Geller F, Gudbjartsson D, Sigurdsson H, Jonsdottir T, Hrafnkelsson J, Johannsson J, Sveinsson T, Myrdal G, Grimsson HN, Jonsson T, von Holst S, Werelius B, Margolis S, Lindblom A, Mayordomo JI, Haiman CA, Kiemeney LA, Johannsson OT, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K (2007) Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet* **39**: 865–869.

Stacey SN, Manolescu A, Sulem P, Thorlacius S, Gudjonsson SA, Jonsson GF, Jakobsdottir M, Bergthorsson JT, Gudmundsson J, Aben KK, Strobbe LJ, Swinkels DW, van Engelenburg KC, Henderson BE, Kolonel LN, Le Marchand L, Millastre E, Andres R, Saez B, Lambea J, Godino J, Polo E, Tres A, Picelli S, Rantala J, Margolin S, Jonsson T, Sigurdsson H, Jonsdottir T, Hrafnkelsson J, Johannsson J, Sveinsson T, Myrdal G, Grimsson HN, Sveinsdottir SG, Alexiusdottir K, Saemundsdottir J, Sigurdsson A, Kostic J, Gudmundsson L, Kristjansson K, Masson G, Fackenthal JD, Adebamowo C, Ogundiran T, Olopade OI, Haiman CA, Lindblom A, Mayordomo JI, Kiemeney LA, Gulcher JR, Rafnar T, Thorsteinsdottir U, Johannsson OT, Kong A, Stefansson K (2008) Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet* **40**: 703–706.

Swift M, Reitnauer PJ, Morrell D, Chase CL (1987) Breast and other cancers in families with ataxia-telangiectasia. *N Engl J Med* **316**: 1289–1294.

The Breast Cancer Association Consortium (2006) Commonly studied single-nucleotide polymorphisms and breast cancer: results from the Breast Cancer Association Consortium. *J Natl Cancer Inst* **98**: 1382–1396.

Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, Hankinson SE, Hutchinson A, Wang Z, Yu K, Chatterjee N, Garcia-Closas M, Gonzalez-Bosquet J, Prokunina-Olsson L, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Diver R, Prentice R, Jackson R, Kooperberg C, Chlebowski R, Lissowska J, Peplonska B, Brinton LA, Sigurdson A, Doody M, Bhatti P, Alexander BH, Buring J, Lee IM, Vatten LJ, Hveem K, Kumle M, Hayes RB, Tucker M, Gerhard DS, Fraumeni Jr JF, Hoover RN, Chanock SJ, Hunter DJ (2009) A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat Genet* **41**: 579–584.

Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, Seal S, Ghoussaini M, Hines S, Healey CS, Hughes D, Warren-Perry M, Tapper W, Eccles D, Evans DG, Hooning M, Schutte M, van den Ouweland A, Houlston R, Ross G, Langford C, Pharoah PD, Stratton MR, Dunning AM, Rahman N, Easton DF (2010) Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat Genet* **42**: 504–507.

Udler MS, Meyer KB, Pooley KA, Karlins E, Struewing JP, Zhang J, Doody DR, MacArthur S, Tyrer J, Pharoah PD, Luben R, Bernstein L, Kolonel LN, Henderson BE, Le Marchand L, Ursin G, Press MF, Brennan P, Sangrajrang S, Gaborieau V, Odefrey F, Shen CY, Wu PE, Wang HC, Kang D, Yoo KY, Noh DY, Ahn SH, Ponder BA, Haiman CA, Malone KE, Dunning AM, Ostrander EA, Easton DF (2009) FGFR2 variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation. *Hum Mol Genet* **18**: 1692–1703.

Zheng W, Long J, Gao YT, Li C, Zheng Y, Xiang YB, Wen W, Levy S, Deming SL, Haines JL, Gu K, Fair AM, Cai Q, Lu W, Shu XO (2009) Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat Genet* **41**: 324–328.

Supplementary Information accompanies this paper on British Journal of Cancer website (http://www.nature.com/bjc)