Computational Methods to Study Gene Regulation Using Genomic, Epigenomic and Chromosome Conformation Data.

BY

DAMIAN ROQUEIRO

B.S., Universidad Tecnológica Nacional, Buenos Aires, Argentina, 1997
M.S., University of Illinois at Chicago, Chicago, 2003

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Bioinformatics
in the Graduate College of the
University of Illinois at Chicago, 2013

Chicago, Illinois

Defense Committee:

Yang Dai, Chair and Advisor
Jie Liang
Tanya Berger-Wolf, Computer Science
Amy Kenter, Microbiology and Immunology
Asgerally Fazleabas, Michigan State University

To my parents, for your love and support.

To my wonderful wife, for your dedication and unwavering love.

ACKNOWLEDGMENTS

I am extremely grateful to the members of my committee for their precious time and knowledge. A special thanks to the committee chair, Dr. Yang Dai, whose academic expertise, patience and kindness have guided me during my years as a PhD student. As my advisor, Dr. Dai has been instrumental in helping me challenge my understanding of research topics, always striving to obtain simple, elegant and mathematically sound solutions to complex problems. Thank you Dr. Dai for all your help and encouragement throughout these arduous years.

Thank you Dr. Jie Liang, Dr. Tanya Berger-Wolf, Dr. Amy Kenter and Dr. Asgerally Fazleabas for agreeing to serve on my committee. I have enjoyed immensely working with you, and I hope to be fortunate enough to have the chance of working with you again in the future. I have learned valuable lessons from our discussions, from attending the classes you taught and, most importantly, from the way you conduct yourself as professionals and researchers.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

TABLE OF CONTENTS (Continued)

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF TABLES (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

FIGURE **PAGE**

LIST OF ABBREVIATIONS

3'RR	3' regulatory region
BN	Bayesian network
bp	Base pair(s)
CDF	Chip description file
CPT	Conditional probability table
ChIP-Seq	Chromatin immunoprecipitation followed by sequencing
CpG	Cytosine-guanine dinucleotides
DAG	Directed acyclic graph
DBN	Dynamic Bayesian network
DNA	Deoxyribonucleic acid
FASTQ	FASTA sequence and its Quality scores
FDR	False discovery rate
KEGG	Kyoto encyclopedia of genes and genomes
me-mRNA-pipe	Methylation-mRNA analysis pipeline
NGS	Next-generation sequencing
ORF	Open reading frame

LIST OF ABBREVIATIONS (Continued)

PAM	Partition around medoids
PE	Paired-end
PFM	Position frequency matrix
PWM	Position weight matrix
P/M	Present or Marginal (status of a probeset)
RF	Random forest
RMA	Robust multichip averaging algorithm
RNA	Ribonucleic acid
TFBS	Transcription factor bnding site
TF	Transcription Factor
TSS	Transcription start site
UPGMA	Unweighted Pair Group Method with Arithmetic mean
UTR	Untranslated region

SUMMARY

Transcriptional regulation in eukaryotes is the process in which different cells regulate the expression of genes. It is extremely complex and the adequate regulation of genes at precise times is what makes many cellular processes viable. Additionally, errors or disruptions in the transcriptional machinery can often compromise the livelihood of the cell or cause disease. In the past few years, novel genomic techniques have been developed to probe the regulatory mechanisms of genes. These techniques include next-generation sequencing, for example, to determine the exact location of DNA-bound regulatory proteins and sophisticated methylation arrays among others. Here we describe a set of computational methods that approach the process of gene regulation from three different research perspectives. Firstly, we explore the standard view of transcription factors binding directly to DNA to promote or repress the expression of genes. The understanding of transcription regulation is enhanced when considering how microRNAs regulate genes at a post-transcriptional phase. Secondly, we analyze how other epigenetic factors, such as DNA methylation, can affect gene expression. Thirdly, we delve into a more complex scenario within the nucleus of the cell where we consider gene regulation as the product, not only of epigenetics or acting transcription factors, but also of the three-dimensional conformation of chromosomes.

The significance of our work is based on the fact that it provides an encompassing view of the complex nature of gene regulation. Because of constant advances in experimental genomics there is a need to develop new analysis methods to cope with the ever increasing volume of biological data that are generated. The deliverables from each of the research aims mentioned above will include, in addition to sound mathematical formulations of how to model the problems, a set of generic (executable) tools from which other researchers can benefit.

CHAPTER 1

INTRODUCTION

1.1 Motivation

This thesis traces how our thoughts and knowledge evolved while exploring mechanisms that regulate gene expression. The organization of this document follows the chronological order of the research questions we have addressed. Each chapter builds on its predecessor by strengthening our understanding of statistical methods and by tackling on more challenging problems. The degree to which a problem is more challenging is not only related to its algorithmic complexity, but to the biological knowledge and the bioinformatics technologies that are required to master it.

We start by exploring algorithms to perform DNA sequence analysis in search of regions with high similarity to a query sequence. These topics are normally covered in Computer Science courses due to their algorithmic nature and potential for optimization. In this context, a simple and mechanistic understanding of how transcription factors work inside the cell will suffice to get a student with Computer Science background –like this humble writer– to get started in bioinformatics. That is precisely why this topic was the first one we embarked on.

Sequence analysis can provide useful predictions of where a transcription factor may bind but, without any biological context, these predictions are too inflexible and possibly unrealistic. Binding of a transcription factor to the promoter of a gene may only occur under specific conditions and it may not always trigger the expression/repression of the gene. As a result of this, and in order to improve our predictions we took into consideration gene expression profiles obtained from microarray experiments. Microarrays, as a technology, have been instrumental in studying how gene expression varies in response to stimuli or in the presence of disease. Therefore, a natural evolution of our work was to take our predictions derived from sequence analysis and square them with true biological signals. In addition to having to master the technology (i.e., microarray analysis), we decided to build a probabilistic framework that would benefit from vast amounts of publicly available expression data. In true machine-learning fashion, we developed a predictor of transcription factors and microRNAs as potential regulators of genes in molecular pathways.

Another biological mechanism that affects gene expression is DNA methylation. DNA methylation is known to regulate genes in a cell, even after cell differentiation. In recent years, new microarrays were developed to detect methylation at a genome-wide scale. These new technologies probe hundreds of thousands of potential methylation sites, resulting in almost a 10-fold increase in volume of data when compared to traditional microarrays. Thus, a natural progression to better understand gene regulation is to combine the analysis of traditional gene microarrays with that of DNA methylation arrays. To that effect, we developed a mathematical model that equates the expression of a gene with the level of methylation at different sites overlapping with the gene. In order to make our model accessible to other researchers, we developed a software product that implements the model and which we intend to make freely available.

Finally, expanding our knowledge about gene regulation will not be possible without a good understanding of the spatial conformation of DNA in the cell. We now know that genes located close to each other in three dimensional space are more likely to share the same expression status: either active or inactive. In the active state, for example, closely co-located genes in 3D space can easily share transcription factors or other elements of the transcription machinery. Likewise, the inactive state of some genes may be the result of a specific DNA conformation that prevents transcription factors from accessing the genes. It is precisely the description of the algorithms we developed to process chromosome conformation data that brings this thesis to its conclusion.

1.2 Thesis outline

Chapter 2 discusses the fundamentals behind the use of sequence analysis to predict binding sites of transcription factors. The first part of the chapter provides a review of the components involved in sequence analysis, finalizing with a description of our contribution and the methods that we developed.

The contents of this chapter are based on the following publications:

- **D. Roqueiro**, J. Frasor and Y. Dai. "bindSDb: A Binding-information Spatial Database". *Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference*, pp.573-578 (2010) doi: 10.1109/BIBMW.2010.5703864.
- P. Yin, **D. Roqueiro**, L. Huang, J.K. Owen, A. Xie, A. Navarro, D. Monsivais, J.S. Coon V, J.J. Kim, Y. Dai, S. E. Bulun. "Genome-Wide Progesterone Receptor Binding: Cell Type-Specific and Shared Mechanisms in T47D Breast Cancer Cells

and Primary Leiomyoma Cells". *PLoS ONE*, 7(1) (2012): e29021. doi:10.1371/journal.pone.0029021.

- W. Mu, **D. Roqueiro**, Y. Dai. "A Local Genetic Algorithm for the Identification of Condition-Specific MicroRNA-Gene Modules". *Scientific World Journal* (2013). doi:10.1155/2013/197406.

In **Chapter 3** we present a probabilistic framework that ranks transcription factors and microRNAs, within specific molecular pathways, based on their effect on gene regulation. Our results show that the framework is useful at providing predictions that are based on disease-specific conditions.

The contents of this chapter are based on the following publications:

- L. Huang, **D. Roqueiro**, Y. Dai. "Analyzing mRNA and microRNA co-expression profiles to identify pathways and their potential regulators in ER+ and ER- breast tumors". *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pp.932-935 (2011) doi: 10.1109/IEMBS.2011.6090210
- **D. Roqueiro**, L. Huang, Y. Dai. "Identifying Transcription Factors and microRNAs as Key Regulators of Pathways Using Bayesian Inference on Known Pathway Structures". *Proteome Science* (2012) doi:10.1186/1477-5956-10-S1-S15.

In **Chapter 4**, when we focus on DNA methylation, we develop a methodology aimed at reducing the dimensionality of the data. Our methodology was implemented as a software product that provides the researcher with a smaller but more reliable set of hypotheses to test in which a group of methylated sites appears to have a strong effect on the expression of a gene. We expect that when our tool is used in the context of a specific disease, it can provide a list of methylated prioritized sites that can be studied as potential biological markers for the disease.

The contents of this chapter are based on the following publication and submission:

- H. Hu, **D. Roqueiro**, Y. Dai. "Prioritizing predicted cis-regulatory elements for co-expressed gene sets based on Lasso regression models". *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pp.6853-6856 (2011) doi:10.1109/IEMBS.2011.6091690.
- **D. Roqueiro**, M.T. Dyson, S.E. Bulun and Y. Dai. "me-mRNA-pipe: A pipeline for the integrative analysis of methylation and mRNA data". In submission.

In **Chapter 5**, we adopt the encompassing view that gene expression, as well as DNA methylation, are affected by the three-dimensional conformation of chromosomes. Our goal is to reliably detect and quantify long-range chromatin interactions in the nucleus. This is a first step towards the ultimate goal of understanding how the 3-dimensional conformation of chromosomes affects the expression of genes. The chapter starts by addressing the methods to transform raw experimental data –representing chromatin interactions– into a coherent and noise-free visual representation of those interactions. The chapter then concludes by presenting a methodology to reliably compare interaction data from two different cells. Although our focus will be on one specific locus dedicated to the production of immunoglobulins, we lay the algorithmic foundations to approach these problems in a generic way and leave the more ambitious goal of correlating DNA interactions with gene expression as my future work.

CHAPTER 2

PREDICTING BINDING SITES OF TRANSCRIPTION FACTORS THROUGH SEQUENCE ANALYSIS.

2.1 Introduction

In Chapter 1, we indicated that transcription factors (TFs) are regulators of gene expression. Simply put, TFs are proteins whose regulating activity is performed by either:

- Binding directly to DNA in the regulatory region of genes.
- Recruiting other proteins to form larger protein complexes, which will ultimately bind to DNA.

Depending on the outcome of their regulatory activity we can classify TFs as *activators* and *repressors*. Generally, activators initiate transcription by interacting with the basal transcription machinery in the promoter of a gene. They can bind to DNA directly or indirectly through another co-activator. Repressors, on the other hand, can negatively affect gene expression by competing with an activator for the same binding location (Lewin et al., 2011).

The previous definition provides a simple, yet powerful, model of how TFs regulate genes. In all fairness, there are many other well-known mechanisms through which activators and repressors affect gene expression. One such mechanism is the recruitment of proteins that modify the conformation of chromatin making it more or less suitable for transcription activity

(Lewin et al., 2011). As a side note, we have developed computational methods to identify changes in chromatin conformation and these methods are the subject of Chapter 5.

In regards to our analysis of TFs, although direct binding to DNA is not necessary for a protein to be considered a transcription factor, the remaining of this chapter will focus on a subset of TFs that are known to bind directly to DNA and whose binding patterns have been experimentally determined. A binding pattern is a representation of the DNA sequence to which a TF is known to bind. These patterns, also known as **motifs**, are computed by analyzing many experimentally obtained transcription factor binding sites (TFBSs) of the same TF.

Motifs are normally modeled using a **position weight matrix** (PWM). We describe what a PWM is and how it is calculated in the next section. Then, the remainder of the chapter focuses on how to use PWMs to identify putative TFBSs in DNA sequences.

2.2 Preliminaries

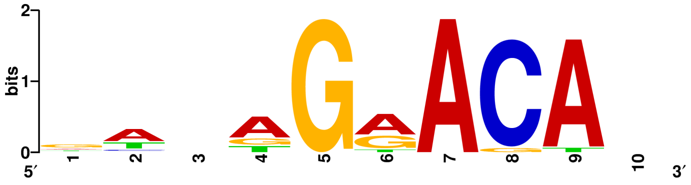
In order to find a probabilistic representation of TFBSs, as it was mentioned before, the PWM of a TF is computed using a set of experimentally confirmed binding sites. After the sites to which the TF binds have been aligned, the next step consists in obtaining a position frequency matrix (PFM). A PFM is a matrix where the rows represent one of four DNA nucleotides: adenine, cytosine, guanine and thymine = {A, C, T, G}, and the columns are the positions of the binding site. The matrix indicates, for row i and column j, the frequency of nucleotide$_i$ at position$_j$. Figure 1(b) shows a PFM for the TF Progesterone receptor derived from 20 experimentally obtained TFBSs shown in Figure 1(a) (source: TRANSFAC ver. 2010.1).

Sequence
GTTTGTACAG
TTAAGAACAG
TTAAGAACAG
GCTAGAACAT
GAGGGGAGAA
AAAAGGACTC
GAGTGAACAG
AAGAGGACAT
GAAAGAACAC
CAAAGGACAG
GAAAGGACAT
AGTAGAACAT
GAGAGGACAT
GTGGGAACAT
GCTGGAACAA
AACGGGACAA
TTTTGAACAC
GACAGAACAC

(a) Binding sites

	1	2	3	4	5	6	7	8	9	10
A	4	11	7	12	0	11	20	0	19	3
C	2	2	2	0	0	0	0	19	0	5
T	10	1	6	4	20	7	0	1	0	6
G	4	6	5	4	0	2	0	0	1	6

(b) PFM obtained from the binding sites



(c) Sequence logo

Figure 1. PFM of progesterone response element, half-site (TRANSFAC Id: V$PR_Q2)

Figure 1(c), known as sequence logo, is a derivative of the PFM and a visual representation of motifs that has been widely adopted in the literature. The heights of nucleotides at each position in the logo are proportional to the sequence conservation of nucleotides at that position. The sequence conservation at position j is measured as (Schneider and Stephens, 1990):

$$R_{seq}(j) = 2 - \left(- \sum_{x \in \{\mathtt{A},\mathtt{C},\mathtt{T},\mathtt{G}\}} \hat{f}(x,j) log_2 \hat{f}(x,j) \right) \tag{2.1}$$

where the first term (number 2) represents the maximum uncertainty $= log_2 4$ because we have 4 nucleotides; $\hat{f}(x,j)$ is the normalized frequency of nucleotide x at position j computed as $\hat{f}(x,j) = \frac{f(x,j)}{N}$ where N is the number of binding sites and $f(x,j)$ is the frequency obtained from the PFM for nucleotide x at position j. Once $R_{seq}(j)$ has been obtained for every j, the height of a letter in the sequence logo is computed as $\hat{f}(x,j)R_{seq}(j)$. The sequence logo in Figure 1(c) was obtained with `WebLogo` (Crooks et al., 2004).

2.2.1 Obtaining a PWM

At this point we have all the information we need to compute a PWM. In simple terms, a PWM characterizes the binding affinity of a transcription factor in the same way the PFM does, except that the frequencies are converted to log-scale.

We proceed by computing an estimate of the corrected probability of seeing nucleotide x at position j (Wasserman and Sandelin, 2004):

$$\hat{p}(x,j) = \frac{f(x,j) + p \cdot C}{N + C} \tag{2.2}$$

where $f(x,j)$ is obtained from the PFM for nucleotide x at position j; p is the background probability of nucleotide x and is assumed to be uniform in all 4 nucleotides (i.e., $p = 0.25$); N is the number of binding sites used to compute the PFM; C is a pseudo-count that corrects for nucleotides not present at certain positions. In our case we use $C = \sqrt{N}$ (Wasserman and Sandelin, 2004).

It is clear from Equation 2.2 that the corrected probability $\hat{p}(x,j)$ is not conceptually different from simply using $p(x,j) = \frac{f(x,j)}{N}$. Finally, the weights in the PWM are computed for each nucleotide x at position j as (Wasserman and Sandelin, 2004):

$$PWM(x,j) = log_2 \frac{\hat{p}(x,j)}{p} \tag{2.3}$$

where, $\hat{p}(x,j)$ is the corrected probability described in Equation 2.2 for x at position j and $p = 0.25$. The equation determines the weight of the nucleotide based on the corrected probability and how much it deviates from the background probability.

The PWM corresponding to the PFM in Figure 1 is shown in Table I. Note, for example, how the matrix assigns negative weights in position 9 to any nucleotide different from `A`. This is because almost all binding sites, as shown in Figure 1(a), have an `A` in that position.

2.2.2 Scoring a sequence with a PWM

The weights in a PWM are used to determine a **similarity score** between a sequence S and the PWM. Say that $S = [s_1, s_2, \ldots, s_n]$ is of length n, where n is the number of columns in the PWM ($n = 10$ in the examples above). The score of S will be the sum of the individual scores of the nucleotides s_j with $j = 1, 2, \ldots, n$. In formal terms, the similarity score of S is defined in Equation 2.4 as (Wasserman and Sandelin, 2004):

TABLE I

PWM COMPUTED FOR HALF-SITE OF PROGESTERONE RESPONSE ELEMENT (TRANSFAC ID: V$PR_Q2)

	1	2	3	4	5	6	7	8	9	10
A	-0.257	0.986	0.408	1.100	-2.452	0.986	1.787	-2.452	1.717	-0.571
C	-0.972	-0.972	-0.972	-2.452	-2.452	-2.452	-2.452	1.717	-2.452	0.000
T	0.862	-1.530	0.218	-0.257	1.787	0.408	-2.452	-1.530	-2.452	0.218
G	-0.257	0.218	0.000	-0.257	-2.452	-0.972	-2.452	-2.452	-1.530	0.218

$$score_S = \sum_{j=1}^{n} \text{PWM}(s_j, j) \tag{2.4}$$

where s_j determines the row in the PWM (if $s_j = \texttt{A}$, then row=1, and so forth) and j is the column.

Because PWMs represent the binding affinity of TFs, they are used to scan DNA sequences in order to predict putative TFBSs. The idea behind this is that if we have a hight similarity score between a short genomic segment and the PWM of TF$_k$, we can make a prediction that TF$_k$ may bind to that genomic segment. We call this a putative TFBS of TF$_k$ because we do not have any experimental validation that TF$_k$ binds to that location.

2.2.3 Motivation to use PWMs

There are hundreds of PWMs stored in private and publicly available databases. An example of the former is TRANSFAC (Matys et al., 2006) and, of the latter, is JASPAR (Bryne et al., 2008). The trove of data in these databases has spurred the development of many algorithms that attempt to find TFBSs which are over-represented in sets of DNA sequences. The idea behind this kind of evaluation derives from the analysis of gene expression microarrays in which sets of genes that have similar expression profiles are assumed to be co-regulated. In order to make sense of these sets of co-regulated genes, a common strategy is to analyze their promoter regions hoping to identify if some TFs may bind to them and, as a consequence, be responsible of their regulation. There are many tools that attempt to predict TFBSs in DNA sequences using PWMs and they are described below.

The Transcription Element Listening System (TELiS) (Cole et al., 2005) detects transcription-factor binding motifs (TFBMs) that are over-represented in the promoter region of a group of genes submitted as a query by the user. This is achieved by differentiating the genes as up- and down-regulated and by utilizing a pre-compiled set of matrices. These matrices store the frequency with which each specific TFBM is detected in each promoter. Different sets of matrices are obtained when the underlying algorithm (MatInspector) (Quandt et al., 1995) is invoked at three different stringency levels of .80, .90, and .95. The definitions of the TFBMs used in TELiS were obtained from JASPAR and from the original free version of TRANSFAC v3.2 (Wingender et al., 1996).

Another system recently developed is SMART (Systematic Motif Analysis Retrieval Tool) (Veerla et al., 2010). SMART downloads RefSeq annotations from the UCSC Genome Browser (Fujita et al., 2010) into a local MySQL database in addition to the DNA sequences of promoter regions. It then proceeds to scan the sequences looking for TFBSs using matrices from

TRANSFAC. This produces TFBS/promoter databases that contain all genes annotated in the UCSC Genome Browser in addition to the location of the predicted TFBS.

Other similar tools include oPOSSUM (Ho Sui et al., 2007), PAP (Chang et al., 2007), TOUCAN2 (Aerts et al., 2005) and the Genomatix suite (Cartharius et al., 2005). The results provided by these tools, including the results from SMART, are ultimately used to identify clusters of TFBSs that co-occur in a set of DNA sequences.

As useful as the above mentioned tools are, they report groups of TFs found to be over-represented in a set of DNA sequences. Here we want to tackle a different problem. The question we want to address is: Given a PWM representing one TF, can we assign a *measure of confidence* to a similarity score of a single (putative) TFBS?

2.3 Predicting TFBSs in the promoter regions of genes

If you recall from the previous section, we described how to compute a similarity score between a PWM and a short DNA sequence (of the same width as the PWM). What we are proposing now is to scan the entire promoter region of a gene with a PWM, and to report *hits* for the locations where the similarity scores are high. This, in turn, implies that the TF may bind at those locations. The main questions we will address in this section are:

- What similarity score can be used as a threshold to report a *hit*?
- Can we assign a measure of confidence, in the form of a p-value, to a similarity score?

If the width of the PWM is w, we will scan the promoter using a sliding window of width w with a step of 1 bp. Figure 2 depicts this process.

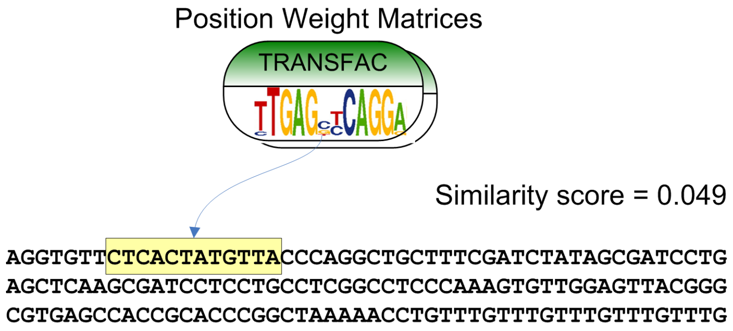

Position Weight Matrices
TRANSFAC
Similarity score = 0.049
AGGTGTTCTCACTATGTTACCCAGGCTGCTTTCGATCTATAGCGATCCTG
AGCTCAAGCGATCCTCCTGCCTCGGCCTCCCAAAGTGTTGGAGTTACGGG
CGTGAGCCACCGCACCCGGCTAAAAACCTGTTTGTTTGTTTGTTTGTTTG

Figure 2. Using a sliding window to scan a DNA sequence with a PWM.

For each substring of length w in the promoter, we will obtain a similarity score. The issue at hand is: which of these scores are high enough to be considered hits? Our goal in predicting a TFBS is to assign a measure of statistical significance to a similarity score. Thus, we can report hits when the scores have a statistically significant p-value (say, $p \leq 0.01$). In essence, we are shifting the problem of finding a “cold” threshold (for the similarity score) to the problem of finding a “warm” threshold based on a p-value and to which we can assign some probabilistic interpretation.

Finding a p-value for the score of a PWM in a sequence has proven to be an elusive problem. In fact, it has been shown to be NP-hard (Touzet and Varré, 2007). As a result of this, many databases and tools simply rely on score-thresholds given to the weighted matching algorithm of choice. For example: TELiS, SMART and oPOSSUM require a score-threshold input by the user. In other scenarios (Thijs et al., 2004), a p-value for a match can be found when comparing the similarity scores to the appropriate background. And this leads us to our next question: What is an appropriate background?

2.3.1 Background model

In addition to the tools described at the end of section 2.2.3, there is a different category of algorithms that also attempt to find over-represented TFBSs in a set of DNA sequences but do not rely on pre-compiled PWMs. They are called *de novo* motif discovery methods because they attempt to discover hidden motifs in the sequences. What is interesting about these methods is that they are able to assign p-values to the discovered motifs by modelling them against a random background. Therefore, an analysis of what types of background are used by these methods seems pertinent and will help us identify an appropriate background for our problem.

MEME (Bailey and Elkan, 1994) is a very sophisticated and widely used algorithm that uses as random background a single nucleotide frequency distribution computed from the input sequences. In contrast, BioProspector (Liu et al., 2001) models random DNA sequences with Markov models of different order (0 to 3). These higher order background models have been shown to yield better motif discovery results (Thijs et al., 2001). Moreover, algorithms with seemingly unrelated applications, such as gene discovery, rely on Markov models of higher order to identify genes in DNA sequences. One such example is Glimmer 2.0 (Delcher et al., 1999) which uses up to 8th-order Markov chains to identify microbial genes.

But what exactly is a kth-order Markov chain? And how is it applied to a DNA sequence?

2.3.1.1 Markov chains

Using a Markov chain as a model to generate DNA sequences implies that the probability of a nucleotide at position i of the sequence depends solely on the nucleotides at positions $1 \ldots i-1$. The number of positions before i that we need to explore determines the **order** of the Markov chain. Formally, if we have a DNA sequence $S = [s_1, s_2, \ldots, s_n]$ where each nucleotide $s_i \in \{\texttt{A}, \texttt{C}, \texttt{T}, \texttt{G}\}$, a kth-order Markov chain is defined as:

$$P(x_n = s_n | x_{n-1} = s_{n-1}, \ldots, x_1 = s_1) = P(x_n = s_n | x_{n-1} = s_{n-1}, \ldots, x_{n-k} = s_{n-k}) \qquad (2.5)$$

where x_i represents the nucleotide at position i and s_i is the *label* of the nucleotide. We can see in Equation 2.5 that the value of the nucleotide at position n only depends on its k predecessors.

In the case when $k = 0$ we assume nucleotides are independent of each other and their occurrence in a sequence S is determined by their frequency in S. On the other hand, when $k = 1$ we are asserting that $P(x_i = s_i | x_{i-1} = s_{i-1})$ and the nucleotide at position i depends only on the previous position. This dependency is captured in the form of a **transition probability**. Given the fact that we have 4 nucleotides, and we can transition from any one of them to the others, we will need 16 transition probabilities to model a DNA sequence with a Markov chain of order 1. Figure 3 illustrates such a model.

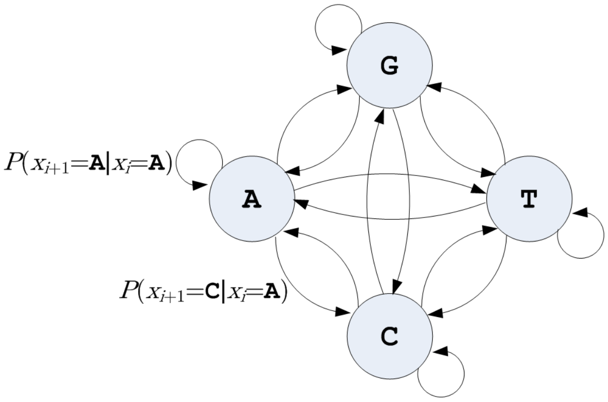


Figure 3. Markov chain of order 1 for DNA

2.3.2 Creating random DNA sequences with a Markov model of order 2

For our random background, we decided to use a Markov chain of order 2. In particular, we created a model for each promoter of a gene. For gene k, the model was created in the following way:

1. Obtain the genomic coordinates of the transcription start site (TSS) of gene k.
2. Obtain the DNA sequence of the region flanked by 1 Kb before and 1 Kb after the TSS of gene k. If the gene is in the negative strand, reverse complement the sequence.
3. Analyze the nucleotide composition of the sequence and create a background model M_k using a Markov process of order 2.
4. Use M_k to create 100 random sequences of the same length as the original one.

The reason why we chose to create a model M_k for each gene k, is the result of analyzing the promoter regions of all genes in the human genome (±1 Kb from the TSS). We took a random sample of 2,144 genes and for each promoter region we determined the frequency of the 16 possible dinucleotides: `AA`, `AC`, `AT`,..., `GG`. Under the assumption of a uniform random distribution, we would expect to see all frequencies $= \frac{1}{16} = 0.0625$. But as it can be seen in the boxplot at the top of Figure 4, there is a large divergence between the frequencies. Many of them have a median value (center line in the boxes) close to 0.0625, but others have a much larger/lower median. The minimum and maximum frequencies are 0.1% and 27.06% for `CG` and `CC` respectively.

The bottom part of Figure 4 shows similar results but from a different perspective: rows are genes and the columns are the frequencies of each dinucleotide in the promoter of the gene. The rows were rearranged using hierarchical clustering, and the dendogram in the left of the heatmap shows how the genes cluster at different levels. It is interesting to see that there are two main clusters: a) top part of the heatmap, with genes whose promoters have larger concentrations of `CC`, `CG`, `GC` and `GG`, and b) the bottom part where the frequencies of these dinucleotides is less than expected by chance.

The `CG` dinucleotides are also referred to as CpGs where the "p" is a phosphodiester bond and they are present with a very low frequency in a substantial portion of the genes analyzed. This is an important result because long stretches of CpGs –known as **CpG islands**– are known to be associated to 72% of gene promoters (Saxonov et al., 2006). If you believe there is an ambiguity when referring to *long stretches of CpGs* it is because there is no objective measure

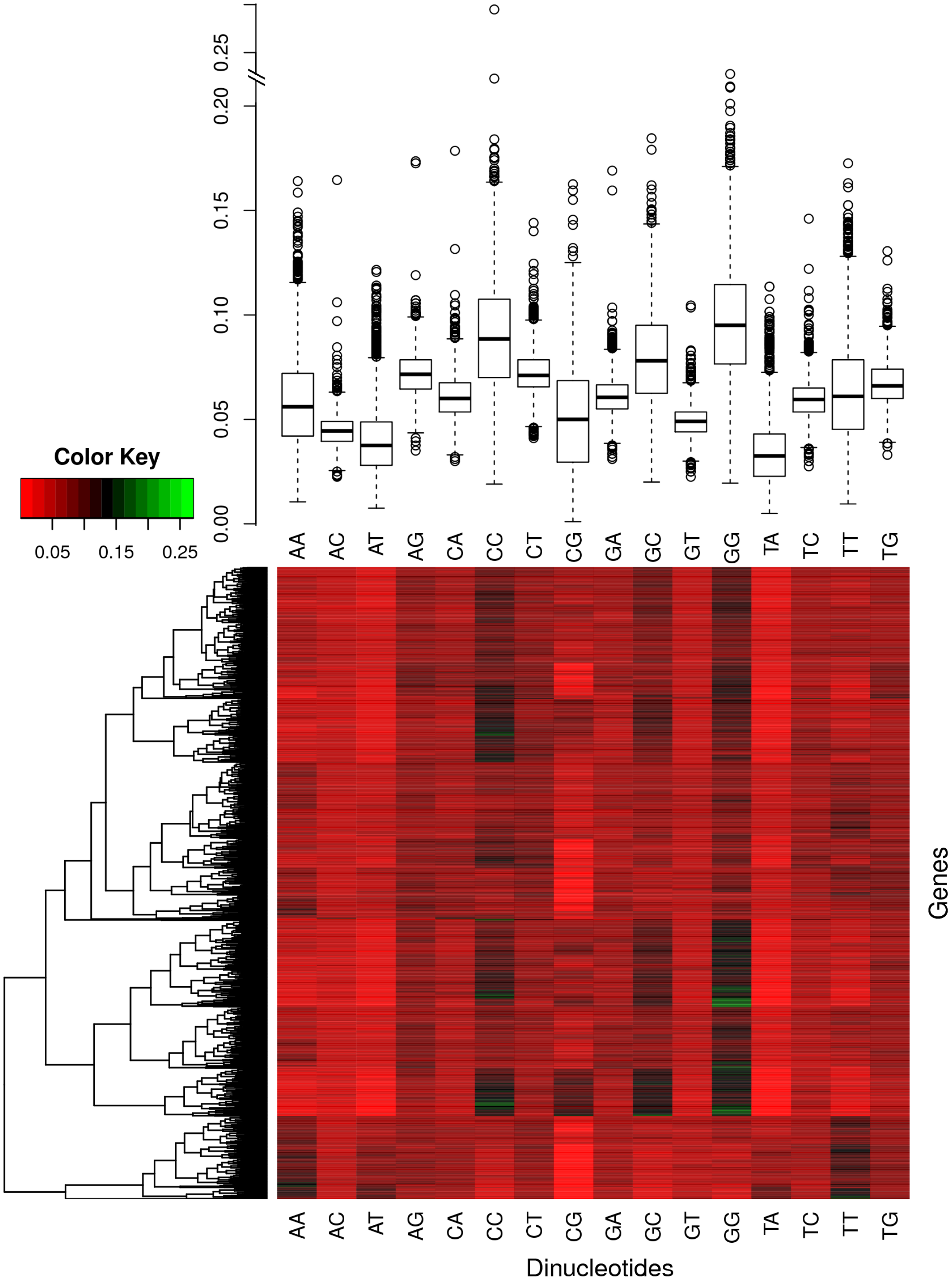


Figure 4. Distribution of dinucleotides in promoters of a random set of 2,144 genes. (top) Boxplot with overall distributions; (bottom) Heatmap with hierarchical clustering, each row of the heatmap is the promoter of a gene.

for what constitutes a CpG island. This has been a source of controversy in the literature and here we wanted to accentuate its ambiguous nature. Setting that aside, CpG islands have received notoriety because they are known to resist DNA methylation –the addition of a methyl group to a cytosine– and DNA methylation has strong effects in gene expression (Deaton and Bird, 2011). The importance of CpG islands and methylation in the promoter of genes is the topic of Chapter 4. There we describe a methodology we developed that attempts to show the value in deviating from the classical approach to methylation analysis, focusing on other areas of genes and not just the promoter.

The entire analysis described above was performed with genomic information downloaded from the UCSC Genome Browser (Kent et al., 2002; Fujita et al., 2010), RefSeq track for the Feb. 2009 (hg19) human assembly. Of a total of 35,954 RefSeq genes, there were 26,517 different TSSs. This is because many genes have different alternative splicings but share the same TSS. For each of these TSSs we executed the steps mentioned at the beginning of this section.

The output of those steps was a model M_k for each gene k. A graphical representation of a Markov model of order 2 would be convoluted and not as easy to understand as the model depicted in Figure 3. On the other hand, the transition probabilities are easy to represent in matrix form and Table II shows the matrix obtained from the promoter of the gene KCTD17 potassium channel tetramerization domain containing 17 (RefSeq Id: NM_024681).

From the table we see that, for example, the probability of seeing a `C` after the dinucleotide `GC` is 0.431 and each row in the transition matrix adds up to 1. After the model M_k is derived using the promoter of gene k, we use it to generate random sequences against which we can

TABLE II

TRANSITION MATRIX CORRESPONDING TO MARKOV MODEL OF ORDER 2 (REFSEQ ID: NM_024681)

	A	C	T	G
AA	0.324	0.190	0.143	0.343
AC	0.236	0.340	0.255	0.170
AT	0.188	0.219	0.172	0.422
AG	0.224	0.259	0.150	0.367
CA	0.174	0.273	0.174	0.380
CC	0.192	0.344	0.277	0.188
CT	0.089	0.357	0.204	0.350
CG	0.157	0.306	0.083	0.455
TA	0.257	0.270	0.125	0.349
TC	0.171	0.335	0.220	0.274
TG	0.136	0.333	0.136	0.395
TT	0.280	0.236	0.108	0.376
GA	0.250	0.273	0.205	0.273
GC	0.192	0.431	0.254	0.123
GT	0.103	0.485	0.206	0.206
GG	0.234	0.234	0.172	0.359

evaluate the p-value of similarity scores of different PWMs. This is discussed in detail in the next section.

2.4 Proposed methods

2.4.1 Computing p-values for similarity scores

As we mentioned before, exact computation of p-values for similarity scores obtained using PWMs is NP-hard (Touzet and Varré, 2007). For a PWM of width w, a naïve approach will enumerate 4^w possible sequences of width w, obtain a similarity score for each of them, and then compare the distribution of these scores to the score for which we want to compute the p-value.

Because exhaustive enumeration is unfeasible for large w, several approximation algorithms have been developed. A tree-based approach (using tries) (da Fonseca et al., 2008) extends an algorithm based on a branch-and-bound method to obtain p-values (Bejerano, 2003). It avoids full enumeration by setting bounds to word prefixes and, therefore, efficiently computes p-values. A different approach (Touzet and Varré, 2007), utilizes discretized score distributions and an iterative approach at estimating the p-values. An implementation of this algorithm is publicly available [1].

In our work, we will take a different approach at approximating p-values. Before describing the steps to compute the p-values, let's summarize all the data we have at our disposal:

- DNA sequence S_k corresponding to the promoter of gene k (± 1 Kb around the TSS). Length of S_k is 2 Kb.
- M_k: Markov model of order 2 computed from S_k.
- 100 random DNA sequences R_k^i with $i = 1, \ldots, 100$ created using the model M_k. All R_k^i have the same length as S_k.
- PWM_q representing the binding affinity of TF_q. Width of PWM_q is w.
- A similarity score t computed with PWM_q for a subsequence of $S_k = \{s_m, s_{m+1}, \ldots, s_{m+w-1}\}$.

An empirical p-value is computed in the following way:

[1] URL: `http://bioinfo.lifl.fr/TFM/TFMpvalue/`

1. Sample N words without replacement, of length w from R_k^i, $i = 1, \ldots, 100$.

2. For each sampled word j, use PWM$_q$ to compute a similarity score t_j

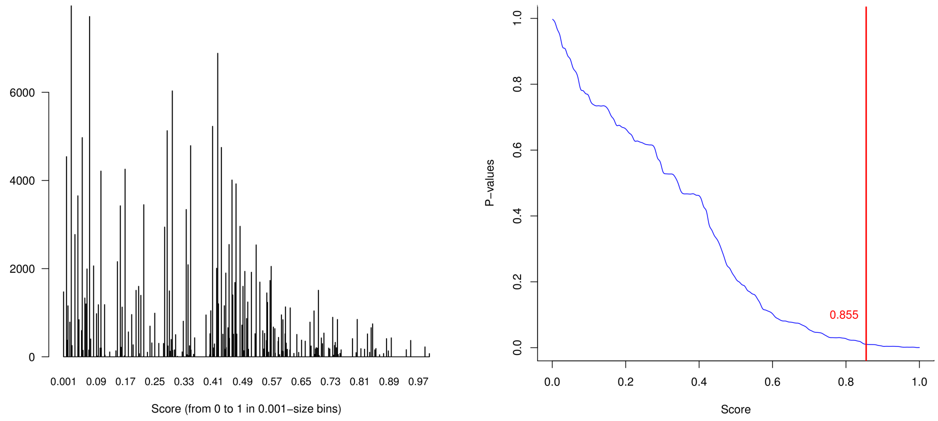
3. Compute the p-value for t as:

$$p_t = \frac{\sum_{j=1}^{N} \mathbf{1}(t_j \geq t)}{N} \tag{2.6}$$

where $\mathbf{1}(\cdot)$ is the indicator function.

Because of the random nature of the sequences created using the M_k model of each promoter, the same similarity score t computed with the same PWM$_q$ may have different p-values in different promoters. In other words, the set of scores with a p-value less than, say, 0.01 may differ between promoters. An example of this is illustrated in Figure 5 and Figure 6. Both figures show the distribution of similarity scores –and their respective p-values– of the PWM V$PAX4_02 corresponding to the TF PAX4 (Paired box gene 4). The difference between the figures is that the distribution of scores were obtained from two different promoters. Figure 5 was obtained from the promoter of the gene FUT8 (Fucosyltransferase 8; RefSeq Id: NM_004480) whereas Figure 6 was obtained from the promoter of THY1 (Thy-1 cell surface antigen; RefSeq Id: NM_006288). This clearly shows that different promoters, due to their nucleotide composition, give rise to different similarity score distributions and, thus, different p-value distributions.

Additionally, the figures mark with a red line the similarity score threshold with a p-value of 0.01. Because of the positive skew in Figure 5(a), we reach a p-value of 0.01 with a smaller similarity score (score of 0.855 in FUT8 vs. 0.989 in THY1). This distinction is important because if we were to apply the same threshold to all genes we will have no control over the false positive rate at predicting TFBSs.



(a) Histogram of similarity scores (bin size = 0.001)

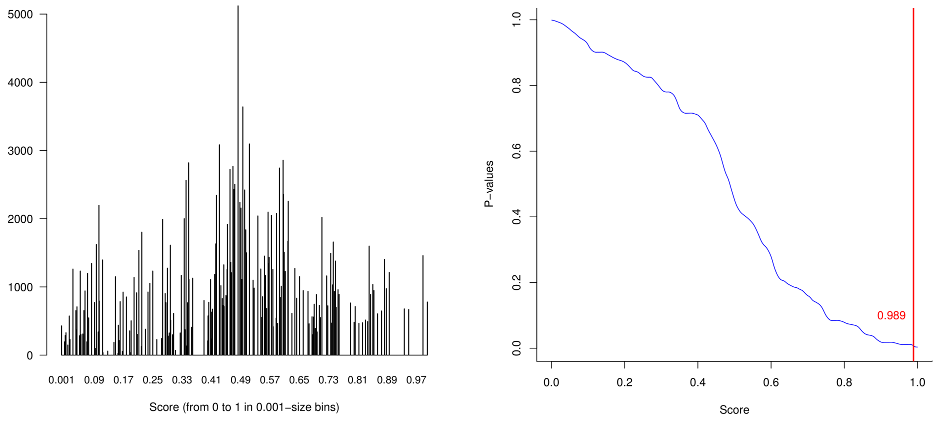
(b) P-values assigned to (bins of) similarity scores

Figure 5. Similarity scores and *p*-values for PAX4 (TRANSFAC Id: V$PAX4_02), in the promoter of the gene FUT8.

To compute all the similarity scores mentioned above we used Match (Kel et al., 2003) as our weighted matching algorithm.

2.4.2 Predicting TFBSs in ChIP-Seq data

Let me start by stating that I am aware the title of this section may resemble an oxymoron. The reason of the apparent contradiction is that **chromatin immunoprecipitation followed**



(a) Histogram of similarity scores (bin size = 0.001)

(b) P-values assigned to (bins of) similarity scores

Figure 6. Similarity scores and p-values for PAX4 (TRANSFAC Id: V$PAX4_02), in the promoter of the gene THY1.

by sequencing (ChIP-Seq) has become the de facto technology to determine binding sites of TFs (ENCODE, 2012). ChIP-Seq maps at a genome-wide level in vivo interactions between DNA-binding proteins, including TFs. Because of its high-throughput nature, ChIP-Seq can provide the location of hundreds of thousands of TFBSs at an approximate 100-bp resolution. The method can be summarized in the following way: a) the chromatin immunoprecipitation component (ChIP) rescues DNA sites where a specific TF is bound using an antibody for the TF; b) these DNA sites are then sequenced (the Seq component in ChIP-Seq), c) mapped to a reference genome and d) their enrichment is quantified (Johnson et al., 2007).

This begs the following question: "What TFBSs are we trying to predict in experimentally confirmed TFBSs?" In this section we address the issue of predicting the binding sites of TF_k

in ChIP-Seq data obtained for TF$_q$. Our goal is to predict which TF$_k$ may be co-located in the experimentally confirmed binding sites of another TF$_q$. Co-location is important in TFs because the transcriptional machinery normally requires a multitude of TFs to be present before transcription can take place. Additionally, we need to make a distinction between **direct binding** and **indirect binding** of a TF. ChIP-Seq will report sites where a TF binds directly to DNA as well as other sites where the TF binds to another protein or TF which in turn binds to DNA (indirect binding). The latter case will also imply co-location and the rest of the chapter details the method we propose to predict co-location of TFs based on ChIP-Seq data.

2.4.3 Background model

In section 2.3 we described a background model for gene promoters based on a Markov chain of order 2. Here we want to create a different background model to determine if the enrichment of a TF in ChIP-Seq data is statistically significant. We formally state the problem as:

- We have a set of DNA sequences $\mathbf{S} = \{S_1, S_2, \ldots, S_N\}$ where each S_i represents a binding site, reported by ChIP-Seq, of TF$_q$. We do not know if these binding sites represent a direct or indirect binding of TF$_q$. In addition to having the DNA sequences, we have the genomic coordinates $(start, end)$ of each S_i.
- Create a background set of DNA sequences $\mathbf{B} = \{B_1, B_2, \ldots, B_N\}$ where each B_i has the same length as S_i, for $i = 1 \ldots N$.
- With a PWM representing TF$_k$ scan the original sequences in $\mathbf{S}$ and the background $\mathbf{B}$ searching for putative TFBSs of TF$_k$.

- Determine if there is a statistically significant enrichment of putative TFBSs of TF_k in one set over the other.

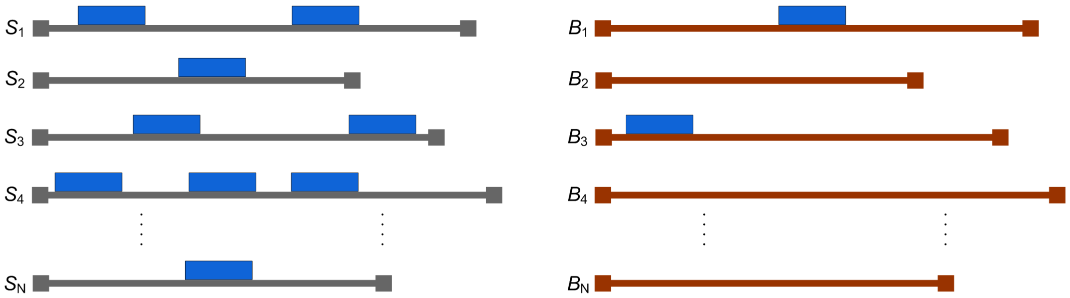
Because $\mathbf{S}$ are true binding sites of TF_q, if we find that a large proportion of these sites have putative TFBFs of TF_k, but TF_k is only present in a small proportion of the background sites in $\mathbf{B}$, we may conclude that TF_k is statistically enriched in $\mathbf{S}$ compared to $\mathbf{B}$. As a result of this, we will predict that TF_k is co-located with TF_q. Figure 7 illustrates this analysis with an example of enrichment of TF_k in $\mathbf{S}$.

The background model for this problem consists of true DNA segments obtained from the vicinity of each S_i (within a random distance of 1 to 10 kb up- or down-stream and not overlapping with any S_j where $j = 1 \ldots i-1, i+1, \ldots N$. This background set is ultimately used to determine a p-value of the enrichment of putative TFBSs of TF_k.

2.4.4 Computing p-values to determine enrichment

Each PWM from TRANSFAC is then scanned in $\mathbf{S}$ and $\mathbf{B}$ using the Match algorithm and a set of similarity score thresholds provided by TRANSFAC named “Minimize False Positive”. These thresholds, as computed by the manufacturer for each PWM, indicate the scores upon which a *hit* may be called. The choice of using true DNA sequences for our background set, as opposed to synthetic ones, was based on our goal to have a better model of binding affinity for PWMs.

For each PWM, the proportion of nucleotides from the true binding sites that contained a reported hit of that PWM was obtained and compared with the proportion of nucleotides in background sites that also contained a hit. For a large sample size, the central limit theorem



(a) Predicted TFBSs of TF$_k$ in true binding sites of TF$_q$

(b) Predicted TFBSs of TF$_k$ in random background

Figure 7. Enrichment of predicted TFBSs of TF$_k$ (blue rectangles) between true binding sites of TF$_q$ (grey lines depicting DNA sequences obtained from ChIP-Seq) and the random background (brown lines, of the same length as the true binding sites).

indicates that the distribution of the sample proportion is approximately normally distributed. Therefore, a z-score was computed for a PWM according to:

$$z = \frac{\hat{p}_t - \hat{p}_b}{\sqrt{\frac{\hat{p}_b(1-\hat{p}_b)}{L}}} \tag{2.7}$$

where $\hat{p}_t$ is the proportion of nucleotides from true binding sites in which the PWM scored a hit; $\hat{p}_b$ is the same proportion but for nucleotides in the background sites; and L is the total number of nucleotides ($L = \sum_{i=1}^{N} length(S_i)$ where S_i is interchangeable with B_i as they have the same length).

The proportion of nucleotides ($\hat{p}_t$ or $\hat{p}_b$) is determined based on the width w of the PWM and the number of hits reported for it. For example, if $N = 10$ and all S_i have a length of 100 nucleotides, then we have $L = 1,000$; assume PWM$_k$ has a width $w = 20$ and 30 non-overlapping hits are reported in $\mathbf{S}$ using Match. Then, the proportion $\hat{p}_t = \frac{30 \cdot w}{L} = \frac{30 \cdot 20}{1{,}000} = 0.6$

Based on the z-score of each PWM, a p-value is computed and later adjusted for multiple hypotheses testing using the Bonferroni correction.

2.5 Conclusion

This chapter presented the principles of sequence analysis and it provided an overview of the necessary elements to compute a similarity score between a PWM and a DNA sequence.

We presented a method to compute p-values based on similarity scores obtained in the promoter region of a gene. A background model that followed a Markov chain of order 2 was created for each promoter. This enabled us to compute p-values for different similarity scores, giving the scores a much needed measure of confidence.

We concluded our analysis by computing p-values for similarity scores obtained from ChIP-Seq data. Here our goal was to determine if a TF$_k$ may be co-located in the experimentally confirmed binding sites of another TF$_q$.

These methods are a good starting point at predicting binding sites for TFs in a single gene, in contrast to the methods described in the introduction of this chapter that found over-represented TFs in a group of genes. Nevertheless, our predictions lack certain flexibility and are not sustained by biological data. This means, for example, that if we predict TF$_k$ binding to the promoter of gene g, two of the main questions that may arise are:

- Will TF$_k$ bind to the promoter of gene g in all tissues?

- Or most importantly, even if TF_k binds to the promoter of gene g, can we quantify the effect this binding has in regulating gene g?

The next chapter addresses these questions and attempts to improve the quality of our sequenced-based predictions.

CHAPTER 3

BAYESIAN INFERENCE TO IDENTIFY TRANSCRIPTION FACTORS AND MICRORNAS AS REGULATORS OF PATHWAYS.

3.1 Introduction

In the previous chapter we introduced algorithmic approaches to predict transcription factor binding sites (TFBSs) based solely on sequence analysis. We used position weight matrices (PWMs) from a licensed release of TRANSFAC ver. 2010.1 (Matys et al., 2006). Additionally, appropriate random backgrounds were created to assign measures of statistical significance to similarity scores obtained from scanning DNA sequences with PWMs.

Here we pick up where we left in the previous chapter, in particular, when it comes to the predictions of TFBSs in the promoter of genes. Despite the fact that we took precautions to compute p-values for our predictions –or even using thresholds set by the manufacturer that claim to minimize false positives– we cannot ignore the fact that these are sequenced-based predictions. As such, some of these predictions may constitute false positives.

Our goal is to integrate the information of these putative TFBSs with expression data obtained from multiple microarray studies. We will focus on a narrow subset of transcription factors (TFs) for which we can find a relationship between the expression levels of the TF and the expression levels of the genes it targets. Our methodology is based on Bayesian inference, particularly, it uses a Bayesian network as a probabilistic framework to describe causal relationships. In addition to the predictions we obtained for TFs from the previous chapter, we will also include in our model predictions of genes as potential targets of microRNAs. TFs and microRNAs are well-known regulators of gene expression. The former bind directly to the regulatory regions of genes in the nucleus whereas the latter regulate the expression of genes at a post-transcriptional phase in the cytoplasm. Although they have different mechanisms of regulation, evidence suggests that TFs and microRNAs regulate target genes in a coordinated way (Martinez and Walhout, 2009).

To determine if a microRNA may target a gene, the state of the art algorithms search for partial complementarity at the 3' untranslated region (UTR) of the gene (Krek et al., 2005; Friedman et al., 2009). In essence, these methods are sequence-based, although they do not rely on PWMs as discussed in Chapter 2.

Our methodology will integrate TF and microRNA target predictions in the context of a disease like breast cancer for which there is an abundance of publicly available expression data. To that effect, we use mRNA and microRNA expression data generated from eight breast tumor studies (Boersma et al., 2008; Desmedt et al., 2008; Miller et al., 2005; Minn et al., 2005; Sotiriou et al., 2006; Wang et al., 2005; Buffa et al., 2011; Enerly et al., 2011). The patients in these studies were divided into two groups: estrogen receptor positive (ER+) and estrogen receptor negative (ER-). ER+ and ER- tumors display different molecular patterns in terms of cell differentiation, proliferation, survival, invasion and angiogenesis. Understanding the distinct molecular mechanisms in tumors with different ER status will provide insight into potential novel targets for breast cancer treatment (Osborne and Schiff, 2011).

It is important to note that our framework is generic enough in that in can be fed multiple datasets of mRNA and microRNA expression data as long as two different phenotypes are available.

3.2 Preliminaries

In order to facilitate the elucidation of the regulatory mechanisms of TFs and microRNAs, several databases have been released based on the analysis of sequence information. Backes et al. (Backes et al., 2010) have compiled a dictionary on microRNAs and their putative pathways based on the enrichment of the predicted microRNA targets for each pathway in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2010). MIR@NT@N (Le Bechec et al., 2011) is a database that stores predicted interactions between: a) a TF and its target genes (including microRNAs) and b) microRNAs and their predicted target genes. These databases facilitate the retrieval of regulatory interactions based on a query list as input but the expression data of mRNA and microRNA are not effectively explored. The analysis tool mirConnX (Huang et al., 2011a) allows the input of concurrent microRNA and mRNA profiling data for an integrative analysis. The targets of TFs and microRNAs are selected based on the association strength between the regulator and its target.

In all the above mentioned work, the analysis of the interactions is focused solely on direct targets. In this project we propose a novel integrative method to analyze microRNA and mRNA expression data in conjunction with sequence-based predicted regulators and the structures of existing molecular pathways. We combine all this information into Bayesian networks, which allow the prediction of pathway regulators, not only based on direct targets but also by inference of the most probable effect of the regulators on other downstream genes.

Bayesian networks (Pearl, 1985) have been extensively used for the reconstruction of gene networks based on microarray expression data. In this context the goal was the inference of interactions and statistical dependencies among genes. These dependencies were, in turn, used to learn the dynamic structure of a regulatory network (Friedman et al., 2000). This methodology has been the foundation for numerous algorithmic approaches. In all these cases, the Bayesian network (BN) –or its more generic dynamic counterpart (DBN)– were used as tools to reverse engineer the gene network, i.e., the interactions between genes were inferred from observational data.

3.3 Proposed method

In this work we do not focus on the task of learning the structure of the BN from expression data. Our goal is to use a known network structure, describing interactions between genes and proteins, for Bayesian inference. The network structure can be any experimentally confirmed interaction network (for example, pathways obtained from KEGG (Kanehisa et al., 2010) or from the Pathway Interaction Database (Schaefer et al., 2009)). Due to the fact that only some TFs and no microRNAs are included in the above mentioned pathways, we must extend the pathways to contain TFs and microRNAs that are predicted to target nodes in the pathway. We further compute conditional probabilities between the nodes in the extended network using expression data, with the ultimate goal of building a BN for each individual pathway. Finally, these BNs receive as evidence a list of differentially expressed genes and provide as output a ranked list of TFs and microRNAs that best explain the expression level of genes in the network. As a result of this, the output TFs and microRNAs are hypothesized to be putative regulators of the pathway.

We started by discretizing the expression data of mRNAs and microRNAs from ER+ and ER- tumor microarray profiles. We subsequently obtained the known structure of 34 KEGG pathways and pre-processed them to guarantee that: a) there were no cycles and b) all nodes in the pathway had expression data. For nodes that passed the pre-processing step we proceeded to obtain lists of TFs and microRNAs that are predicted to target the nodes. We then ranked the TFs and microRNAs based on their ability to predict the expression level of a target gene. We obtained one ranking list per gene and expanded the pathways to include the top 5 TFs and the top 3 microRNAs for each gene in a pathway. Finally, a BN was created for each extended pathway. Inference was performed by entering, as evidence, the statuses (discrete values) of differentially expressed genes in the pathway. The inference process was performed twice with evidence derived for one phenotype and later with evidence derived from the other phenotype. The marginal probabilities were approximated for all unobserved nodes. From these, the TFs or microRNAs with the largest difference in marginal probabilities between phenotypes were considered the most probable regulators of expression in the pathway. An overview of the entire methodology is illustrated in Figure 8.

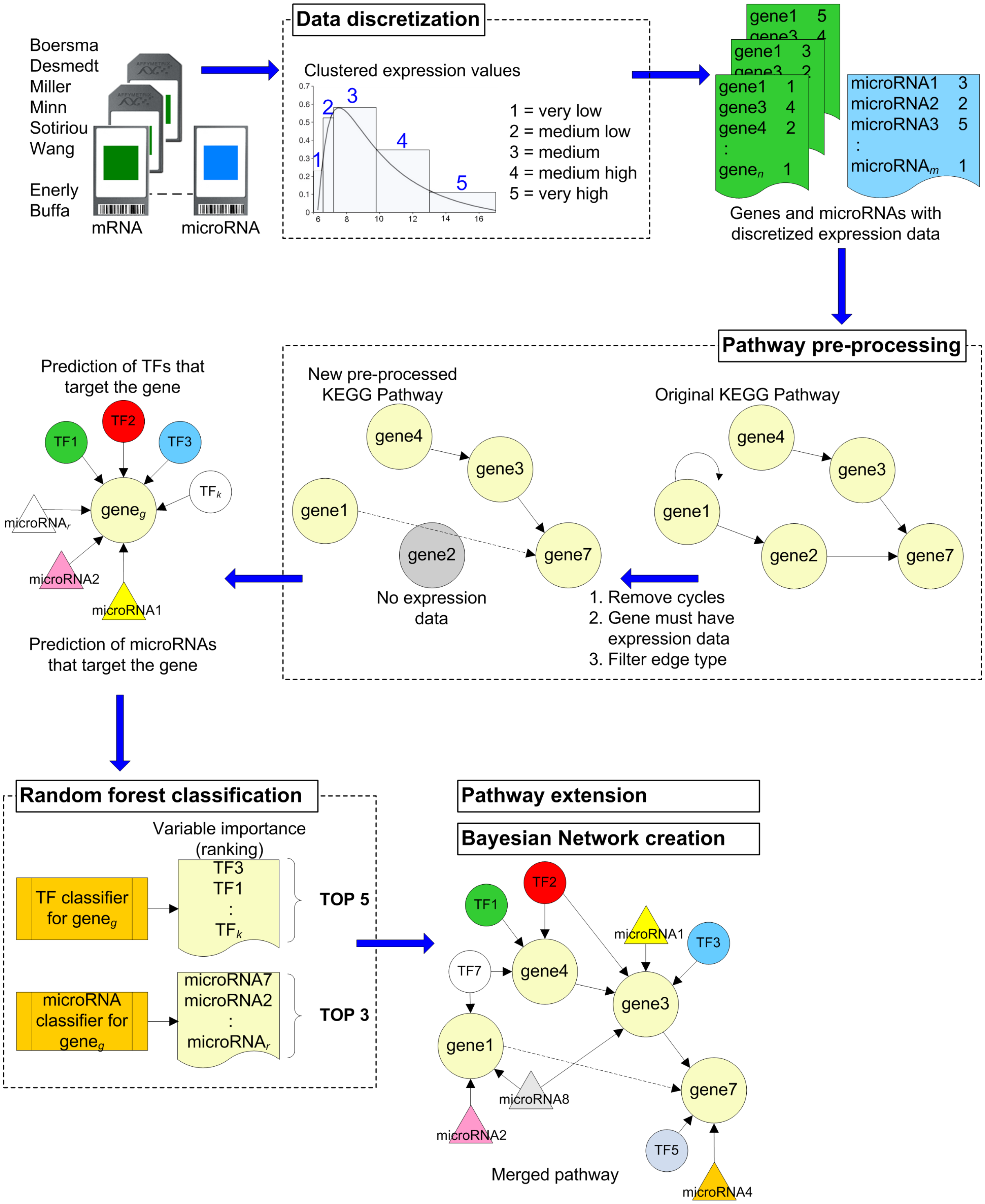


Figure 8. Flowchart of the methodology.

3.3.1 Pre-processing of raw microarray data and data discretization

The raw data from eight studies of estrogen receptor positive (ER+) and estrogen receptor negative (ER-) breast tumors (Boersma et al., 2008; Desmedt et al., 2008; Miller et al., 2005; Minn et al., 2005; Sotiriou et al., 2006; Wang et al., 2005; Buffa et al., 2011; Enerly et al., 2011) were downloaded from the Gene Expression Omnibus (GEO) (Edgar et al., 2002). Table III provides details about the source of the data and the number of samples for each type of tumor. The first six studies contain only mRNA expression profiles whereas the last two (Buffa and Enerly) have concurrent mRNA and microRNA expression profiles on ER+/ER- breast tumors. Herein, we will refer to the datasets using the name provided in Table III.

TABLE III

ANALYZED ER+/ER- EXPRESSION DATASETS

			Number of samples	
Name		Source	ER+	ER-
Boersma	(mRNA)	GSE5847 (Boersma et al., 2008)	41	52
Desmedt	(mRNA)	GSE7390 (Desmedt et al., 2008)	107	51
Miller	(mRNA)	GSE3494 (Miller et al., 2005)	213	34
Minn	(mRNA)	GSE2603 (Minn et al., 2005)	57	42
Sotiriou	(mRNA)	GSE2990 (Sotiriou et al., 2006)	74	24
Wang	(mRNA)	GSE2034 (Wang et al., 2005)	209	77
Buffa	(mRNA)	GSE22219 (Buffa et al., 2011)	122	79
Buffa	(microRNA)	GSE22216 (Buffa et al., 2011)	122	79
Enerly	(mRNA)	GSE19783 (Enerly et al., 2011)	60	35
Enerly	(microRNA)	GSE19536 (Enerly et al., 2011)	60	35

One of the important aspects of our data integration methodology is the fact that different data sources are supported. The case study presented in this chapter illustrates the use of microarray platforms from different manufacturers. Table IV provides details of these platforms.

TABLE IV

PLATFORM DETAILS OF ER+/ER- EXPRESSION DATASETS

Dataset	Platform	Probe mapping
Boersma Desmedt Miller Minn Sotiriou Wang	Affymetrix Human Genome U133A Array (HG-U133A)	Custom CDF
Buffa	Illumina HumanRef-8 v1.0 expression beadchip	Manufacturer's annotation
Buffa (miR)	Illumina Human v1 MicroRNA expression beadchip	
Enerly	Agilent Whole Human Genome Microarray, 4×44K	
Enerly (miR)	Agilent Human miRNA Microarray (V2) Kit, 8×15K	

A custom Chip Description File (CDF) was used to find a unique mapping between probesets and Entrez IDs in the first six datasets. The Affymetrix microarrays used in these studies relied, at the time of their creation, on probes defined upon earlier genome and transcriptome annotations. These annotations differ significantly from our current knowledge, therefore for our analysis we used a custom CDF that conforms to the latest genome and transcriptome available information: custom CDF `HGU133A_Hs_ENTREZG` version 13.0 released on July 2010 (Dai et al., 2005). For the Agilent and Illumina microarrays, their proprietary annotation files were used.

3.3.2 Data normalization

Data normalization of each dataset forced its microarrays to have the same empirical distribution of intensities. The Robust Multichip Averaging algorithm (RMA) (Irizarry et al., 2003) with quantile normalization was used for normalization of the Affymetrix microarrays. Additionally, to minimize the noise level in the subsequent task of data discretization, Affymetrix detection calls were used only for Affymetrix data to identify probesets with low or no level of expression.

The raw microRNA data from Enerly were normalized with the RMA algorithm using the AgiMicroRna package (Lopez-Romero, 2011). The mRNA data in the Enerly study as well as the mRNA and microRNA data in the Buffa study were already normalized by the authors.

There was a very large variability in the Affymetrix microarrays. This can be seen in Figure 9 (first column) where the densities of expression values for the microarrays are plotted as curves. The second column shows the same curves after normalization. The third column shows, for one randomly chosen microarray in the study, how the normalized density curve changes shape when only the expression values of probesets marked as "Present" (P) or "Marginal" (M) are used. This is due to the fact that probesets marked as "Absent" (A) may still have an expression value different from zero. By using P/M calls we followed best practices recommended for Affymetrix microarrays and obtained more reliable data to input to our system.

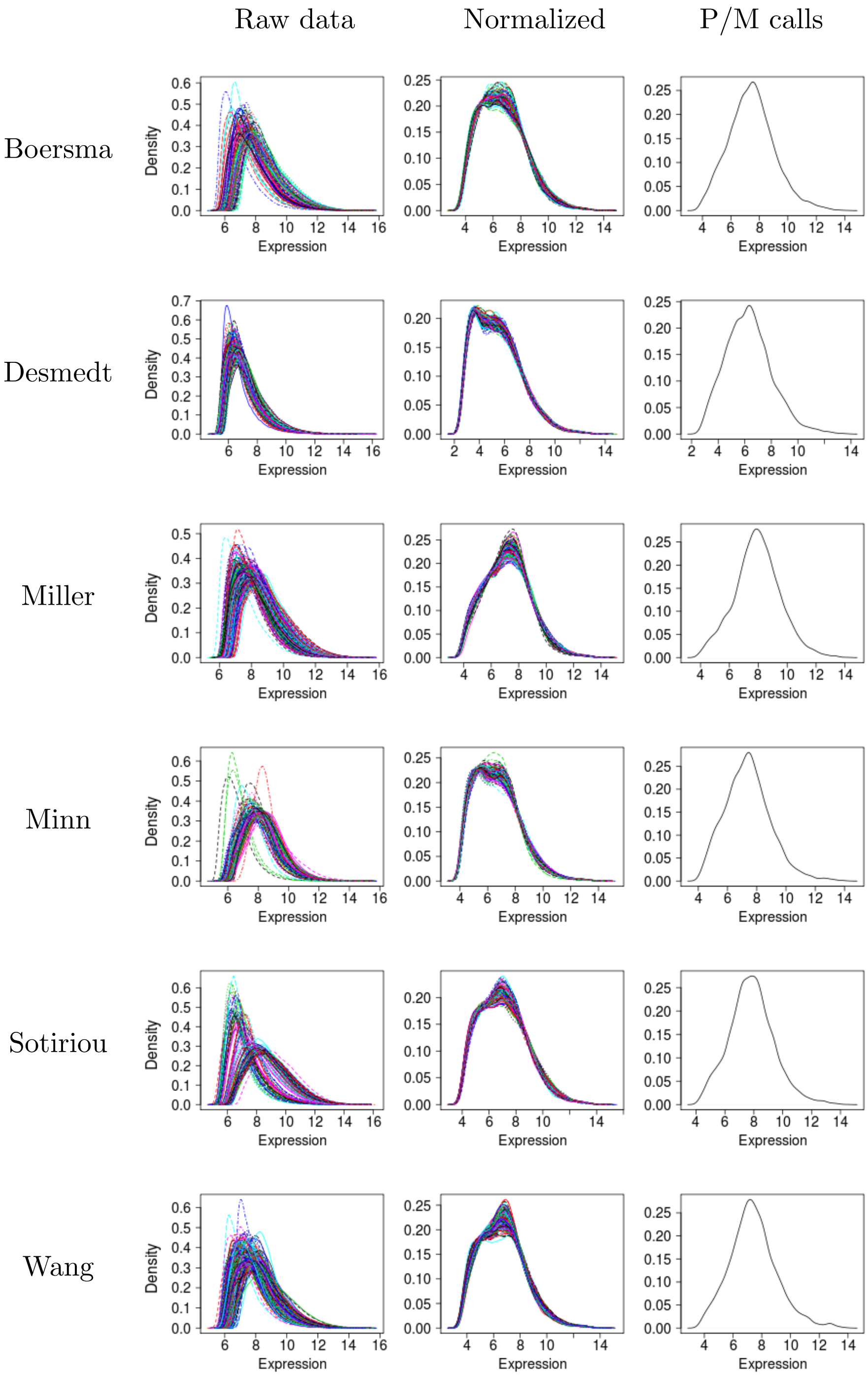


Figure 9. Density of expression data in mRNA datasets.

Figure 10 shows the density curves of the normalized expression values in Buffa and Enerly. The first column of the figure has the mRNA data while the second column has the microRNA data.

Gene expression analysis was performed using R with packages from `Bioconductor` (Gentleman et al., 2004).

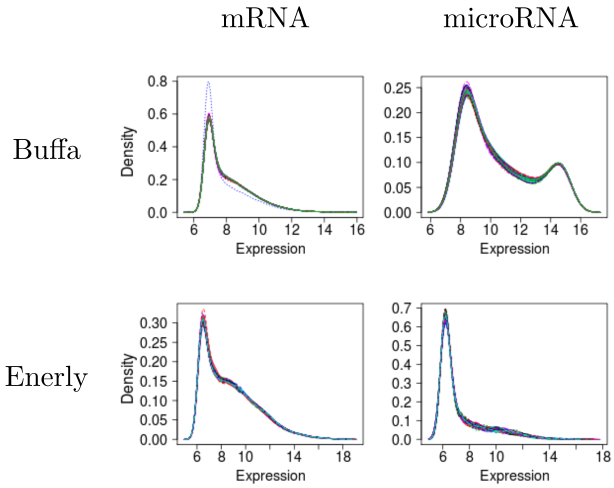


Figure 10. Density of normalized expression data in combined mRNA-microRNA datasets.

3.3.3 Differential expression analysis

Each dataset was analyzed independently to obtain lists of differentially expressed genes between ER+ and ER- samples. These differentially expressed genes were used in two different contexts:

- Differentially expressed genes in the first six datasets were used to determine the most appropriate discretization method. (Data discretization is discussed in Section 3.3.4)
- Differentially expressed genes in the Buffa and Enerly datasets were used as evidence in the Bayesian inference process. (Evidence in the inference process is discussed in Section 3.3.11)

The differential expression analysis was performed on all normalized microarrays. For Affymetrix, a probeset in a study was discarded if it was not marked as “Present” or “Marginal” in more than 85% of the samples, or if the coefficient of variability (CV) of the expression values of the probeset was less than 50% across samples. The `limma` package (Smyth, 2004) with the Benjamini-Hochberg correction for multiple hypothesis testing (Benjamini and Hochberg, 1995) were used for differential expression analysis. The adjusted *p*-value threshold was set to 0.05.

For the Agilent mRNA chips, the normalized expression data were downloaded from GEO and only the probes with unique Entrez IDs were kept. For the Agilent microRNA data, the probes with a detection signal of less than 10% of the samples or not associated with *H.sapiens* were discarded.

The normalized expression data of Illumina mRNA chips were downloaded from GEO and those probes with unique Entrez IDs were retained. Probes with a CV of less than 20% were filtered out. For the Illumina microRNA chips, only probes associated with *H.sapiens* were retained. The differential expression analyses were performed with `limma` as described above.

The remainder of this section provides technical details about the process we followed to obtain differentially expressed genes in each microarray platform. We close this analysis by comparing lists of differentially expressed genes obtained in different datasets to show that they do not overlap well.

For Affymetrix (all mRNA data)

1. Use the Wilcoxon signed rank-based gene expression presence/absence detection algorithm from the `affy` package in R `Bioconductor`. [function `mas5calls`]
2. Discard probesets not marked as "Present" or "Marginal" in 85% of samples.
3. Obtain coefficient of variability (CV) for probesets. If σ is the standard deviation and μ is the mean, $\text{CV} = \frac{\sigma}{\mu}$.
4. Discard probesets with $\text{CV} < 0.5$ across samples
5. Differential expression using `limma` package in R `Bioconductor` [functions `lmFit` and `eBayes`]
6. Adjust p-values using Benjamini-Hochberg correction [function `topTable`]
7. Report probesets with adjusted p-value < 0.05

For Agilent (mRNA)

1. Discard probes that are not associated to an Entrez ID(s)
2. Keep the probe with maximum variance across all samples if a gene has multiple probes
3. Follow steps 5 through 7 for the Affymetrix data

For Agilent (microRNA)

1. Discard probes with detection signal in less than 10% of the samples or not associated with *H.sapiens*
2. Follow steps 5 through 7 of the Affymetrix data

For Illumina

1. Discard probes that are not associated to an Entrez ID(s)
2. Keep the probe with maximum variance across all samples if a gene has multiple probes
3. Discard probes with CV < 0.20 across samples
4. Follow steps 5 through 7 of the Affymetrix data

Table V summarizes the steps mentioned above and shows the different number of probesets that remained after each processing step. The column "Original" lists the number of probeset/probes after removing those not associated to an Entrez ID. The column "Filtering method" indicates how the probes were further filtered.

TABLE V

NUMBER OF MRNA PROBESETS RETAINED AT EACH FILTERING STAGE

	Number of probeset/probes				
Dataset	Original	After P/M calls	Filtering method*	After filtering	Differentially expressed
Boersma	12,133	5,245	1	2,604	690
Desmedt	12,133	6,227	1	3,095	1,672
Miller	12,133	5,932	1	2,952	1,521
Minn	12,133	5,639	1	2,805	1,481
Sotiriou	12,133	6,355	1	3,160	1,341
Wang	12,133	5,753	1	2,859	1,834
Buffa	24,385	–	2,4	12,501	4,723
Buffa (miR)	735	–	2,5	488	150
Enerly	41,094	–	2	17,117	3,808
Enerly (miR)	729	–	3	498	60

*Filtering method (key):

1. Discard probesets with CV < 0.5 across samples
2. For genes with multiple probes, only the probe with maximum variance across all samples was kept.
3. Probes with detection signal in less than 10% of the samples or probes not associated with *H.sapiens* were discarded.
4. Discard probesets with CV < 0.2 across samples.
5. microRNA probes not associated with *H.sapiens* were discarded

3.3.3.1 Overlap among lists of differentially expressed genes

We identified discrepancies when analyzing the lists of differentially expressed genes that were obtained from each dataset. Although the patients in these different studies were divided according to two distinct phenotypes such as ER+ and ER- the differentially expressed genes across studies did not have a large overlap.

For each dataset we ranked the genes according to an ascending order of adjusted *p*-value. We examined the overlap across multiple datasets for all the differentially expressed genes as well as the overlap for the top 100-ranked genes, 200-ranked genes, and other top rankings. Additionally, we wanted to determine if a gene was reported with a different status in two datasets, e.g.: as down-regulated in ER+ samples of one dataset and reported as up-regulated in ER+ of a different dataset. Table VI shows the details.

TABLE VI

OVERLAP OF DIFFERENTIALLY EXPRESSED GENES

	Overlap of 6 Affymetrix datasets	Overlap of Buffa and Enerly mRNA datasets
All genes	150	2,172
Different status	0	4
Ranking < 100	4	36
Ranking < 200	7	70
Ranking < 500	37	197
Ranking < 1,000	116	421
Ranking < 2,000	150	907

It can be seen that only 4 genes overlap when taking the top 100-ranked differentially expressed genes in the first 6 Affymetrix datasets. This was a clear sign, early on, that our integrative analysis could not simply rely on differentially expressed genes. As was mentioned before, for our Bayesian inference we used all genes in the microarrays regardless if they were differentially expressed or not. The information about differentially expressed genes in each of the six Affymetrix datasets was nonetheless useful as it will be illustrated in Section 3.3.4.

When comparing the Buffa and Enerly datasets, there were 907 genes that overlapped among the top 2,000-ranked differentially expressed genes. These 907 genes were used as evidence for the BN inference process.

In these two datasets there are four differentially expressed genes with opposed expression statuses. This means the genes were classified as up-regulated in Buffa and down-regulated in Enerly, or vice versa. Because of their contradictory behavior, these four genes were discarded from our analysis and are detailed in Table VII.

TABLE VII

GENES WITH OPPOSITE DIFFERENTIAL EXPRESSION STATUS IN AGILENT AND ILLUMINA MICROARRAYS

Entrez ID	Gene symbol and description	ER+ status in Buffa	ER+ status in Enerly
10270	AKAP8 A kinase (PRKA) anchor protein 8	up	down
29964	PRICKLE4 prickle homolog 4	down	up
128178	EDARADD EDAR-associated death domain	up	down
374655	ZNF710 zinc finger protein 710	up	down

3.3.4 Discretization of expression data

In a BN, the nodes must have distinct (and finite) discrete states. This required a discretization method to convert the microarray expression data into discrete values to be fed to the BN. We used five states to discretize the expression values of all genes and microRNAs, namely 1=*very low*, 2=*medium low*, 3=*medium*, 4=*medium high* and 5=*very high*.

From Section 3.3.3.1 it became clear that we could not base our integrative methodology solely on differentially expressed genes as they have very little overlap between datasets. We needed to use all the genes and microRNAs for which we were able to obtain expression data. In fact, in order to integrate all eight datasets we had to follow two steps:

1. Determine what genes/microRNAs were common to all eight datasets.
2. Because BNs require discrete states for their nodes, we had to identify a method to convert the expression data obtained from a microarray into discrete values.

3.3.4.1 Genes common to all datasets

The common genes between the first six datasets and the mRNA-Buffa and mRNA-Enerly datasets were kept. The first six datasets, all of them of the same microarray platform, had 12,025 unique Entrez IDs. The mRNA-Buffa and mRNA-Enerly datasets had 15,627 and 17,177 unique genes respectively. A total of **10,722** unique Entrez IDs were common across all mRNA datasets. The microRNA-Buffa and microRNA-Enerly datasets had 510 and 498 microRNAs respectively with **308** microRNAs that overlapped between both datasets.

3.3.4.2 Discretization methods

In order to determine the best discretization method for our microarray data, we analyzed three discretization algorithms:

1. Sigma-mu (σ and μ)
2. Quantile distribution
3. Partition Around Medoids (PAM)

Figure 11 illustrates the overall approach to data discretization. In the figure we see that the expression data of all the M microarrays in a given dataset are discretized one by one.

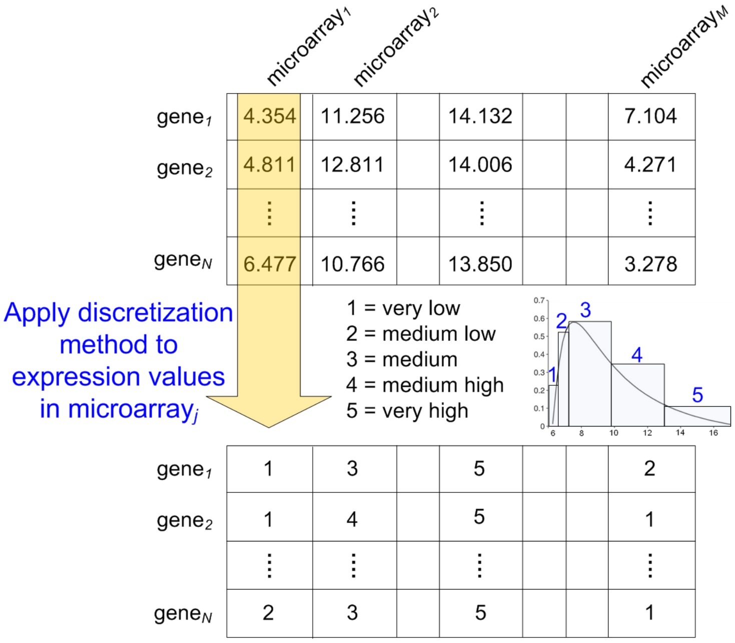


Figure 11. Data discretization overview.

3.3.4.3 Sigma-mu (σ and μ)

This method is based on the mean (μ) and standard deviation (σ) of all the expression values in microarray$_j$. The expression level of a gene/microRNA was compared against how many standard deviations away from the mean it was. The five discrete values were assigned as: *very low* and *very high* ($\geq 2\sigma$ from μ); *medium low* and *medium high* ($\geq 1\sigma$ and $< 2\sigma$ from μ); and medium ($< 1\sigma$ from μ). The pseudocode describing this method is shown in Algorithm 1.

3.3.4.4 Quantile distribution

The density function of the expression values in a microarray was used to obtain estimates of the intervals that accumulated 20%, 40%, 60%, 80% and 100% of the expression values. The sample quantiles were estimated in the following way (Hyndman and Fan, 1996):

1. Sort in ascending order the N expression values in microarray$_j$, $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$

2. To estimate the quantile breakpoint $\hat{Q}_p$ for a given probability p, compute the index k to the list of sorted values as:

   $$\begin{aligned} \hat{k} &= p(N-1)+1 \\ k &= \lfloor \hat{k} \rfloor \end{aligned}$$

   and finally compute the quantile estimate using interpolation

   $$\hat{Q}_p = \boldsymbol{x}_k + (\hat{k} - k)(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)$$

Algorithm 1 Discretization method: Sigma-mu (σ and μ)

```
procedure initializeVariables
    $very\_low \leftarrow 1$                          ▷ Constants for discrete values
    $medium\_low \leftarrow 2$
    $medium \leftarrow 3$
    $medium\_high \leftarrow 4$
    $very\_high \leftarrow 5$
    $\boldsymbol{D} \in \mathbb{N}^{N \times M} \leftarrow 0$                   ▷ Discretized matrix; $N$ genes, $M$ microarrays
end procedure

initializeVariables
for $j = 1 \rightarrow M$ do
    $\boldsymbol{x} \in \mathbb{R}^N \leftarrow$ read expression of $N$ genes in microarray$_j$
    $\mu \leftarrow mean(\boldsymbol{x})$                      ▷ Exclude genes marked as "Absent" from
                                              $mean()$ and $sd()$ computations
    $\sigma \leftarrow sd(\boldsymbol{x})$
    for $i = 1 \rightarrow N$ do
        if $\mu - \sigma \leq \boldsymbol{x}_i \leq \mu + \sigma$ then
            $v \leftarrow medium$                   ▷ One $\sigma$ away from $\mu$
        else if $\mu + \sigma < \boldsymbol{x}_i \leq \mu + 2\sigma$ then
            $v \leftarrow medium\_high$             ▷ Less than $+2\sigma$
        else if $\mu + 2\sigma < \boldsymbol{x}_i$ then
            $v \leftarrow very\_high$               ▷ More than $+2\sigma$
        else if $\mu - 2\sigma \leq \boldsymbol{x}_i < \mu - \sigma$ then
            $v \leftarrow medium\_low$              ▷ Less than $-2\sigma$
        else[$\boldsymbol{x}_i < \mu - 2\sigma$]
            $v \leftarrow very\_low$                ▷ More than $-2\sigma$
        end if
        $\boldsymbol{D}_{ij} \leftarrow v$
    end for
end for
$\boldsymbol{D}_{ij} \leftarrow very\_low$ if gene$_i$ in microarray$_j$ is marked as "Absent"
Use discrete values in $\boldsymbol{D}$
```

The pseudocode describing this process is shown in Algorithm 2. As shown in the algorithm, the discrete value corresponding to a given expression value is the quantile in which the expression value belongs. Quantile 1 [0% to 20%) was assigned a discrete value of *very low.* Quantile 2 [20% to 40%) received a discrete value of *medium low*; all the way to quantile 5 [80% to 100%] which was assigned a discrete value of *very high.* The quantile estimation was implemented using the function `quantile` from the R package `stats`.

3.3.4.5 Partition Around Medoids (PAM)

The expression values of all the genes in microarray$_j$ were clustered into 5 clusters using the Partition Around Medoids (PAM) algorithm. The lowest cluster (1) contained genes whose expression values were in the lowest range of expression level. The remaining clusters, 2 through 5 contained genes with higher expression values than those in the previous cluster. Therefore, genes whose expression values were clustered in the lowest and highest end of the spectrum (clusters 1 and 5) were discretized as *very low* and *very high* respectively. Genes in clusters 2 and 4 were discretized as *medium low* and *medium high.* Genes in the remaining cluster (3) were discretized as *medium.*

For its implementation, we used the function `pam` from the R package `cluster`.

Note on discretization of Affymetrix microarrays: Regardless of the discretization method used, the above processes were applied only to probesets marked as either "Present" or "Marginal" (P/M). Probesets marked as "Absent" were always given a discrete value of *very low.* This is shown in Algorithm 1, lines 12 and 13, where the mean and standard deviations are only computed for genes marked as P/M. Line 29 in Algorithm 1 and line 33 in Algorithm 2 show how genes marked as "Absent" are imputed a discrete value of *very low.*

Algorithm 2 Discretization method: Quantiles

```
function ESTIMATEQUANTILES($\boldsymbol{x}, probs$)
    $N \leftarrow length(\boldsymbol{x})$
    $P \leftarrow length(probs)$
    $\hat{\boldsymbol{Q}} \in \mathbb{R}^P \leftarrow 0$                ▷ Will contain the start and end of each quantile
    for $i = 1 \rightarrow P$ do                 ▷ Iterate for each probability value $p$
        $p \leftarrow probs_i$
        if $p \neq 1.0$ then
            $\hat{k} \leftarrow p \times (N-1) + 1$
            $k \leftarrow \lfloor \hat{k} \rfloor$                ▷ Compute index to array of elements
            $\hat{\boldsymbol{Q}}_i \leftarrow \boldsymbol{x}_k + (\hat{k} - k) \times (\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)$
        else
            $\hat{\boldsymbol{Q}}_i \leftarrow x_N$
        end if
    end for
    return $\hat{\boldsymbol{Q}}$
end function

INITIALIZEVARIABLES                          ▷ Same initialization as in Algorithm 1
$discrete = \{very\_low,\ medium\_low,\ medium,\ medium\_high,\ very\_high\}$
$probs = \{0.0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1.0\}$
for $j = 1 \rightarrow M$ do
    $\boldsymbol{x} \in \mathbb{R}^N \leftarrow$ read expression of $N$ genes in microarray$_j$
    $\boldsymbol{s} \leftarrow sort(\boldsymbol{x})$                ▷ In ascending order
    $\boldsymbol{Q} \leftarrow$ ESTIMATEQUANTILES($\boldsymbol{s}, probs$)
    for $i = 1 \rightarrow N$ do
        $P \leftarrow length(probs)$
        for $k = 1 \rightarrow (P-1)$ do         ▷ Discretize based on the quantile
            if $\boldsymbol{x}_i \geq \boldsymbol{Q}_k$ And $\boldsymbol{x}_i < \boldsymbol{Q}_{k+1}$ then
                $\boldsymbol{D}_{ij} \leftarrow discrete_k$
            end if
        end for
    end for
end for
$\boldsymbol{D}_{ij} \leftarrow very\_low$ if gene$_i$ in microarray$_j$ is marked as “Absent”
Use discrete values in $\boldsymbol{D}$
```

3.3.5 Evaluation of discretization methods

Data discretization has a strong effect over the conditional probabilities assigned to each node in a BN. Therefore, we conducted a comparison of the three discretization algorithms described above (Sections 3.3.4.3 through 3.3.4.5) to determine the one that was most appropriate to our study.

We compared the discrete values obtained from each method to identify the one that created the largest contrast between the two phenotypes in the data (ER+ vs. ER- in this case). To detect this contrast, we used as reference the genes which we had determined to be differentially expressed in each dataset. In theory, if a gene is differentially expressed in a dataset, it means that the expression values of the gene in the ER+ samples are different from the expression values of the same gene in ER- samples. Consider the following example, for dataset$_d$ we have M samples (microarrays). The first s samples correspond to patients categorized as ER+. The remaining $M - s$ samples correspond to ER- patients. Therefore, if gene$_k$ is differentially expressed in dataset$_d$ we can expect the discrete values of gene$_k$ in samples $[1...s]$ to be "significantly different" to the discrete values of the same gene in samples $[s + 1...M]$.

To quantify that difference for a given gene$_k$, we used the unweighted pair-group method arithmetic averages (UPGMA) between ER+ and ER- samples:

$$\Delta gene_k = \frac{1}{|ER+||ER-|} \sum_{x \in ER+} \sum_{y \in ER-} dist(x, y). \tag{3.1}$$

where:

- gene$_k$ must be differentially expressed in dataset$_d$
- ER+ and ER- are the samples for each phenotype
- the distance measure $dist(x, y) = |x - y|$, with x and y being discrete values between 1 and 5.

In summary, all the discrete values of gene$_k$ in ER+ samples were compared against the discrete values of gene$_k$ in ER- samples. This process was repeated for all the differentially expressed genes in dataset$_d$, summing up the $\Delta gene_k$ for all k.

3.3.5.1 Evaluation criteria

Our requirements for a good discretization algorithm were the following:

For differentially expressed genes:

1. The sum of all $\Delta gene_k$ should be large: we wanted a discretization algorithm that maximized the distance between discrete values of different phenotypes.
2. The number of genes that get the same discrete value in both phenotypes should be minimized. For example, we wanted to avoid as much as possible a case where gene$_k$ was given a discrete value of, say *very high*, in all M microarrays of dataset$_d$. This is an important consideration because even if a gene was determined to be differentially expressed in our analysis, the difference in expression values between phenotypes can be subtle enough that the discretization method could potentially generate the same discrete values for both phenotypes.

Conversely, the same requirements (with opposite criteria) were used for genes that were not differentially expressed.

For the rest of the genes (not differentially expressed):

1. The sum of all $\Delta gene_k$ should be low: in this case, it had to be lower than the value obtained for differentially expressed genes.
2. The number of genes with the same discrete value in both phenotypes should be maximized: it is fair to say that if the gene is not differentially expressed, there should not be much variation between phenotypes.

3.3.5.2 Evaluation results

Table VIII contains the results of the metrics mentioned above for differentially expressed genes. The sum of all $\Delta gene_k$ per dataset was divided by the number of genes to obtain an average UPGMA $\overline{\Delta gene_k}$. The column *same* is the count of genes for which the discrete values were the same across all samples of both phenotypes. The column *diff* indicates the count of genes for which at least one discrete value was different between phenotypes. The addition of *same* + *diff*, in each dataset, is equal to the number of differentially expressed genes in that dataset (see Table V in 3.3.3).

When comparing the counts of *same* vs. *diff* obtained for differentially expressed genes we can see that PAM outperformed the other two methods. This was of outmost importance to us because the differentially expressed genes in Buffa and Enerly will later be used as evidence in the BN inference. Figure 12 shows these counts as percentages.

TABLE VIII

DISTANCE METRICS FOR DIFFERENTIALLY EXPRESSED GENES

	PAM			**Sigma-mu**			**Quantile**		
Dataset	$\overline{\Delta gene_k}$	*same*	*diff*	$\overline{\Delta gene_k}$	*same*	*diff*	$\overline{\Delta gene_k}$	*same*	*diff*
Boersma	1.013	0	690	0.373	41	649	0.780	17	673
Desmedt	1.219	0	1,672	0.372	50	1,622	0.811	10	1,662
Miller	0.967	0	1,521	0.356	50	1,471	0.763	29	1,492
Minn	0.929	0	1,481	0.420	60	1,421	0.836	14	1,467
Sotiriou	1.099	0	1,341	0.347	116	1,225	0.718	39	1,302
Wang	0.954	0	1,834	0.353	42	1,792	0.766	23	1,811
Buffa	1.007	1	4,722	0.195	840	3,883	0.528	115	4,608
Enerly	1.098	0	3,808	0.207	1,130	2,678	0.492	338	3,470
Buffa (miR)	0.650	14	136	0.118	58	92	0.313	22	128
Enerly (miR)	0.912	0	60	0.131	36	24	0.469	10	50

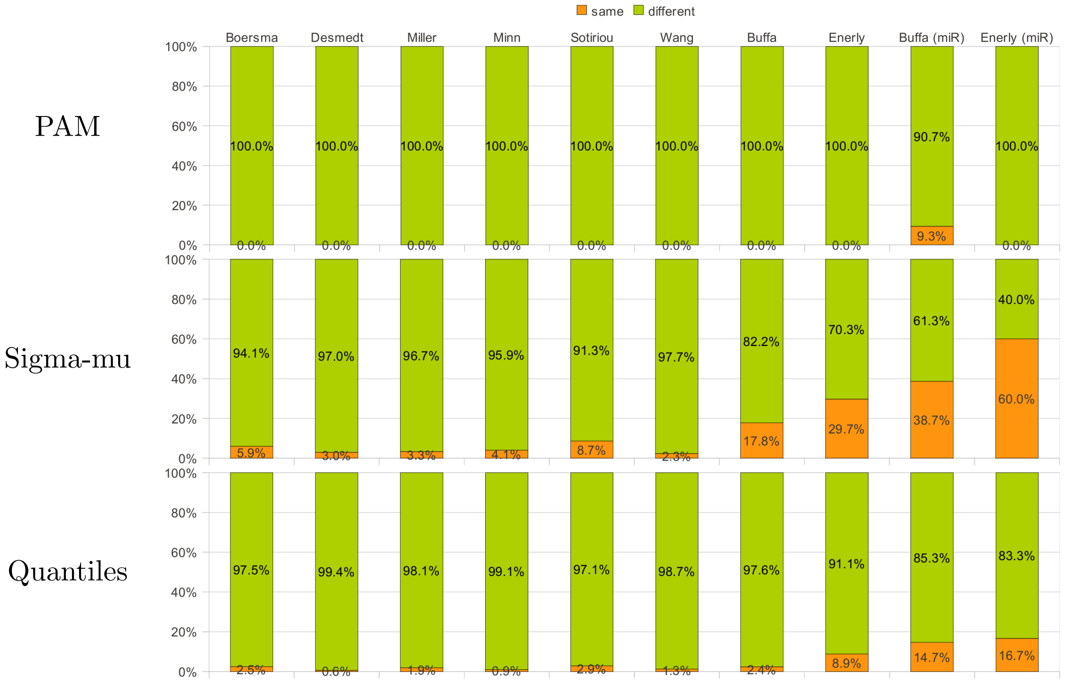


Figure 12. Same vs. different discrete values in differentially expressed genes.

Computing the values $\Delta gene_k$, *same* and *diff* for differentially expressed genes was straightforward. But obtaining these values for the rest of the genes required taking random samples of non-differentially expressed genes, computing the sum of all $\Delta gene_k$, obtaining the counts *same* and *diff* and repeating the process for 100 iterations to average the results. For each dataset the sample size r differed based on the number of differentially expressed genes in it. For example, in the Boersma dataset $r = 690$ (see Table V).

After 100 iterations, the counts were averaged and the results for non-differentially expressed genes can be found in Table IX.

For non-differentially expressed genes, the Quantiles method performed better than the other two. Because the genes were not differentially expressed, we expected to see a higher value for *same* and a smaller value for *diff*. Figure 13 shows the results for the non-differentially expressed genes.

TABLE IX

DISTANCE METRICS FOR NON-DIFFERENTIALLY EXPRESSED GENES

	PAM			**Sigma-mu**			**Quantile**		
Dataset	$\overline{\Delta gene_k}$	*same*	*diff*	$\overline{\Delta gene_k}$	*same*	*diff*	$\overline{\Delta gene_k}$	*same*	*diff*
Boersma	0.570	143.5	546.5	0.326	173.1	516.9	0.386	221.9	468.1
Desmedt	0.700	218.9	1,453.1	0.304	309.6	1,362.4	0.343	478.8	1,193.2
Miller	0.546	260.5	1,260.5	0.283	366.1	1,154.9	0.342	466.4	1,054.6
Minn	0.487	287.7	1,193.3	0.341	287.0	1,194.0	0.382	441.2	1,039.8
Sotiriou	0.690	209.2	1,131.8	0.298	332.4	1,008.6	0.360	416.5	924.5
Wang	0.517	279.6	1,554.4	0.314	349.0	1,485.0	0.352	525.5	1,308.5
Buffa	0.772	0.0	4,723.0	0.167	844.7	3,878.3	0.402	125.3	4,597.7
Enerly	0.858	0.3	3,807.7	0.175	1,156.7	2,651.3	0.346	354.6	3,453.4
Buffa(miR)	0.534	16.5	133.5	0.124	45.1	104.9	0.285	15.5	134.5
Enerly(miR)	0.853	0.1	59.9	0.079	43.5	16.5	0.365	6.4	53.6

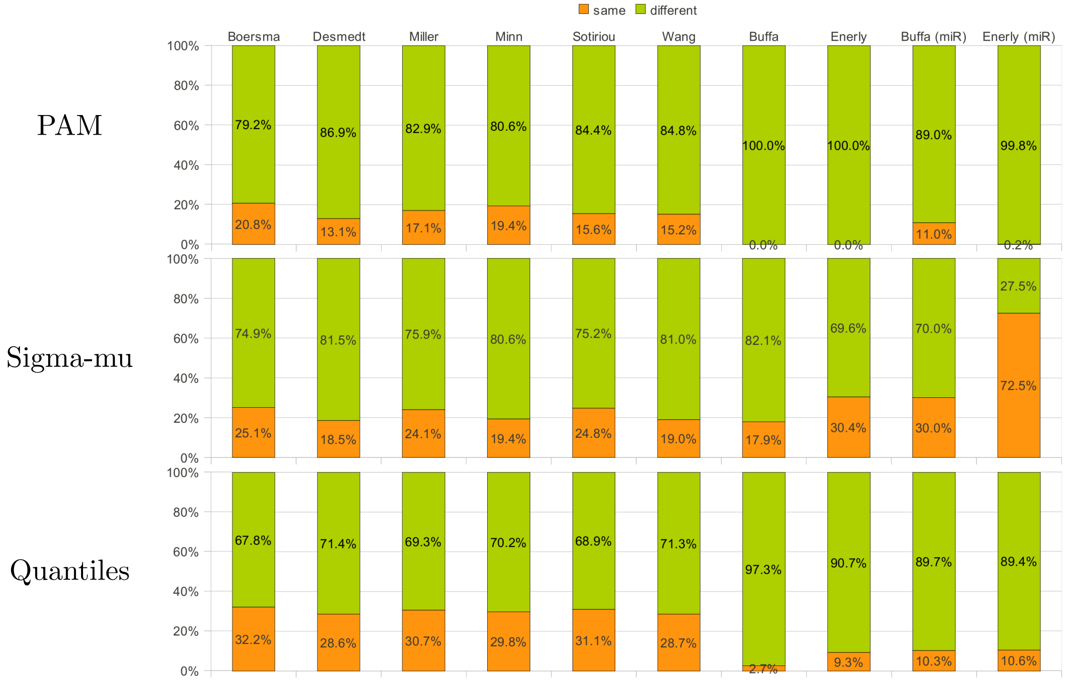


Figure 13. Same vs. different discrete values in randomly chosen non-differentially expressed genes.

Surprisingly, Sigma-mu performed very poorly in both cases. When looking at the shape of the density functions in Figure 9 (third column, normalized values after keeping only P/M probesets), our intuition suggested that this method would be appropriate. But under close examination we can see there is a positive skew in those density functions which, although not as prominent as in Figure 10, it makes this discretization method perform badly in comparison to PAM and Quantiles.

3.3.6 Structure pre-processing of KEGG pathways

The KEGG database provides experimental knowledge in many forms, one of them being molecular networks called KEGG pathway maps. For our work, the pathway maps were analyzed as networks, with directed edges between the nodes representing a known interaction. The pathways analyzed were related to signaling (KEGG Ids 04010-04350) and cancer (05200-05223).

The structure of a pathway including nodes and edges was used as the backbone of a BN. Before the BN could be constructed, a pre-processing step was implemented on the pathway. This pre-processing yielded a new network, based on the original pathway, with the following properties:

- No cycles: The KEGG pathway was transformed into a directed acyclic graph (DAG). Edges that created a loop were discarded.
- Nodes with expression data: The Entrez ID of each node in the pathway was checked against the list of genes that had expression data (10,722 Entrez IDs from our microarray

analysis, see Section 3.3.4.1). Nodes with no expression data were removed. The parents and children (if any) of a removed node were updated to include new edges linking them.

- Limited types of interactions: Only the following interactions annotated in a KEGG pathway were taken into consideration: a) gene expression relations: *expression*, *repression* and *indirect effect*; and b) protein-protein interactions: *activation*, *inhibition* and *indirect effect*.

The package `KEGGgraph` (Zhang and Wiemann, 2009) in `Bioconductor` was used to parse the raw KEGG Markup Language files.

3.3.7 The predicted targets of TFs and microRNAs

Since our goal in implementing a BN for a known pathway is the identification of the set of TFs and microRNAs that are putative regulators of nodes in the pathway, the new network obtained from the previous pre-processing step needed to be expanded to include the TFs and microRNAs that are predicted to target the nodes in the pathway. We followed two different approaches to determine which TFs and microRNAs may target a node in the pre-processed network:

1. **TF target prediction**: `bindSDb` (Roqueiro et al., 2010) is a database we developed to store experimentally proven and predicted transcription factor binding sites. For the prediction portion, the database returns a set of TFs that are predicted to bind to the promoter region of a gene based on sequence analysis. It uses the Match (Kel et al., 2003) algorithm to determine if a TF may bind to the promoter of the gene. Each TF

was represented by one or more position weight matrices from TRANSFAC (ver. 2010.1) (Matys et al., 2006). In our work, for each gene in a pathway, or protein encoded by a gene, we obtained from `bindSDb` all the TFs that are predicted to bind to the promoter region of the gene (in our case defined as ±2 Kb from the transcription start site). Additionally, we obtained from TRANSFAC the information about the genes that encode the predicted TFs (when available). In this way, each gene in the pathway will be associated with a set of genes whose protein products, i.e., TFs, are predicted to target the gene. If one of the predicted TFs was already present in the pathway, then it was not included as a putative regulator of the gene.

2. **microRNA target prediction**: All predictions of microRNAs targeting genes were obtained from the TargetScan Human release 6.0 (Friedman et al., 2009). TargetScan is a microRNA target prediction algorithm that searches highly conserved 3'UTR targets for 8-mer and 7-mer sites matching the seed region of microRNAs. We downloaded target predictions for 677 microRNA families, as defined by TargetScan, and obtained a total of 54,479 unique pairs between microRNA family and target gene.

3.3.8 Pre-selection of TFs and microRNAs

From the previous step we obtain a list of predicted TFs and microRNAs targeting each individual gene in a pathway. Ideally, we would expand the pathways by adding incoming edges to a gene from every TF and microRNA predicted to target the gene. Unfortunately, because of the large number of TFs and microRNAs that may target a gene, this is infeasible. As an example, Table X shows the number of TFs and microRNAs that are predicted to target the genes of three signaling pathways.

TABLE X

NUMBER OF TFS AND MICRORNAS TARGETING GENES IN A PATHWAY

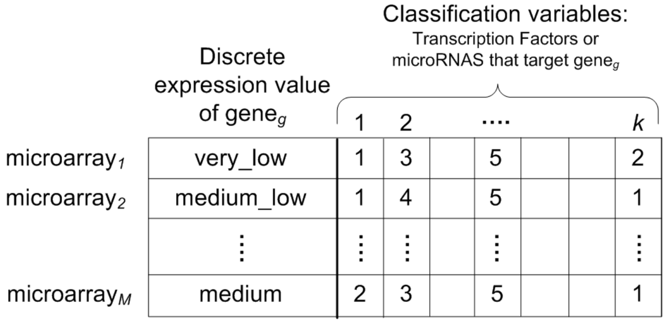
KEGG Id	Pathway name	TFs per node		microRNAs per node	
		Average	*Max.*	*Average*	*Max.*
4010	MAPK signaling pathway	95.3	165	10.7	54
4150	mTOR signaling pathway	90.3	152	17.1	59
4115	p53 signaling pathway	91.8	151	14.2	59

For a complete list of pathways and statistics of the number of nodes targeted by TFs and/or microRNAs, refer to the Appendix, Table XXXIII.

If a node in a BN has more than 100 parents, we simply cannot maintain its conditional probability table (such a table will consist of more than 5^{100} entries). Therefore, it is necessary to limit the number of regulators for each gene. To that respect, we used an improved version of a machine learning approach we previously implemented (Huang et al., 2011b) to obtain a ranking of the TFs and microRNAs that are predicted to target each gene. Based on this ranking, for each gene we chose the top 5 TFs and the top 3 microRNAs and added them to the BN.

Two classifiers were created using the random forest (RF) classification algorithm (Breiman, 2001) on each gene of a pathway. One classifier was based on the expression levels of the associated TFs and, the other, of the microRNAs. For each classifier, the values of the predictor variables were the discretized expression levels of the TFs (mRNAs of the encoding genes) or microRNAs. Our ultimate goal was not to find a classifier to predict the expression level of genes but to use RF to measure the importance of each predictor variable. In this manner, for each gene, a group of TFs and microRNAs that could differentiate the expression level of the gene across different microarrays were obtained.

The layout of the input data to RF is shown in Figure 14. More specifically, the supervised learning predictor for gene$_g$ is defined as $T_g = (y_i, \boldsymbol{x}_i)$ with $i = 1$ to M, where M is the total number of microarrays used in the classifier. For TFs, $M = 980$ (the first six studies listed in Table III) and for microRNAs, $M = 296$ (Buffa and Enerly). The multiclass response vector $\boldsymbol{y}$ contains the M discrete expression levels of gene$_g$ in the microarrays. Each vector $\boldsymbol{x}_i$ has values for k predictor variables (TFs or microRNAs that target gene$_g$), that is, x_{ij} for $j = 1$ to k contains the discrete expression value of predictor j in microarray i. The values were coded according to the data discretization step: from 1 through 5, where 1 = *very low* and 5 = *very high*. For each gene, an ensemble of 2,000 trees (for TFs) and 500 trees (for microRNAs) was created. One third of the variables were randomly chosen at each tree level and one third of the samples were left as out of bag. Variable importance was determined after performing permutations on the trees to assess the change in their predicting power. Each variable was assigned a mean decrease of accuracy score and the ranking of predictor variables for the gene was based on this score. The analysis was implemented with the R package `randomForest`.



Classification variables: Transcription Factors or microRNAS that target gene$_g$

	Discrete expression value of gene$_g$	1	2		....			k
microarray$_1$	very_low	1	3		5			2
microarray$_2$	medium_low	1	4		5			1
	⋮	⋮	⋮		⋮			⋮
microarray$_M$	medium	2	3		5			1

Figure 14. Data layout for random forest classification.

3.3.9 Pathway extension

At this stage, we have all the required information to create a BN for a pathway. The modified pathway obtained after pre-processing in 3.3.6 was extended to accommodate the TFs and microRNAs ranked in Section 3.3.8.

Our RF analysis output two variable importance rankings for each gene: one for the TFs and one for the microRNAs. These rankings list the TFs and microRNAs in decreasing order of the variable importance score assigned to each of them. An extended pathway was then created in the following way:

- For each node g in the pre-processed pathway:

  1. Create nodes representing the top 5 TFs, as reported by RF, that target gene g. Add directed edges from these nodes to g
  2. Create nodes with the top 3 microRNAs targeting g as reported by RF. Add directed edges to g

The top 5 TFs and top 3 microRNAs were only considered if their variable importance score was greater than zero. Also note that the same TF may target more than one gene in the pathway. Therefore, the node for the TF was added just once with multiple edges going from this node to different target genes. The same consideration applied to the microRNAs. This newly merged pathway was then fed to the BN process.

3.3.10 Construction of the Bayesian network

Simply put, a BN can be characterized as (Kwisthout, 2011; Jensen and Nielsen, 2007):

- A directed acyclic graph $G = (V, E)$ where V is a set of variables and E is a set of directed edges between the variables.
- Each variable in V has a finite set of mutually exclusive states.
- For each variable B with parents $A_1, A_2, ..., A_p$ there is a set of parameter probabilities in the form of conditional probability tables (CPTs) that capture $P(B|A_1, A_2, ..., A_p)$.

The first two items have been addressed in previous sections (pre-processing of KEGG pathways and data discretization). The creation of the CPT for a given node in the pathway was implemented in the following way:

1. For nodes with no parents, the CPT was basically a vector representing the prior of the node. It was computed by obtaining the frequencies of each discrete value across all the appropriate microarrays (TFs and genes used the first six datasets of Table III, whereas microRNAs used the Enerly and Buffa datasets).

2. For a node with parents $A_1, A_2, ..., A_p$, the CPT reflected the probability of all possible combinations of states between the node and its parents. The probability of each possible combination was obtained by counting in all appropriate microarrays and then dividing by the total number of observations. A high-dimensional matrix C of 5-by-5-by...(p+1)-times was used to compute the CPT. The matrix C was initialized with 1s to assume that each possible combination of states was possible. Then, for each microarray, the discrete expression values of the node and its parents were obtained as a vector $v = [v_{A_1}, v_{A_2}, ..., v_{A_p}, v_{node}]$. The contents of matrix C at the cell $C[v_{A_1}, v_{A_2}, ..., v_{A_p}, v_{node}]$ were then incremented by one. At the end, each position of C was divided by the sum of all elements in C. The matter of what set of microarrays to use was resolved in the following way:

   - If any of the node's parents $A_1, A_2, ..., A_p$ was a microRNA, the Buffa and Enerly datasets were used
   - Otherwise, the first six datasets listed in Table III were used.

This distinction was absolutely necessary. In order to compute the CPT of a node that had at least one microRNA as parent, we needed to process microarrays that had both expression values for genes/TFs as well as microRNAs. Evidently, because of the number of samples listed in Table III, the CPTs of nodes with a microRNA targeting them were created from fewer observations than nodes whose parents were only TFs or other pathway nodes.

3.3.11 Evidence and inference

An important aspect of a BN is the evidence, i.e., the values assigned to observed nodes. For evidence we used 907 differentially expressed genes between ER+ and ER- samples of the mRNA-Buffa and mRNA-Enerly datasets. These 907 genes were the result of the overlap of the top 2000-ranked differentially expressed genes in each of the two datasets as shown in Table VI. Only those differentially expressed genes that were part of a pathway (not as TFs but as KEGG pathway nodes) were used as evidence.

Once the BN for a pathway was created, we conducted two rounds of inference. The CPTs in our BN were created using all the data from ER+ and ER- samples. Therefore, in order to identify a contrast between these two phenotypes we subjected the same BN to two different sets of evidence corresponding to two scenarios. In scenario #1, the evidence value assigned to a differentially expressed gene was the median of all the discrete values of that gene corresponding to ER+ samples. Conversely, in scenario #2, the evidence was formed by obtaining the medians of the discrete values in ER- samples.

Regardless of which of the two scenarios we are analyzing, for a BN with variables $X_1, X_2, ..., X_{n+s}$ where the evidence $\boldsymbol{e} = [X_{n+1}, X_{n+2}, ..., X_{n+s}]$ and the values of variables $X_1, X_2, ..., X_n$ are unobserved, we would like to obtain $P(X_1, X_2, ..., X_n|\boldsymbol{e})$. This joint probability is defined as:

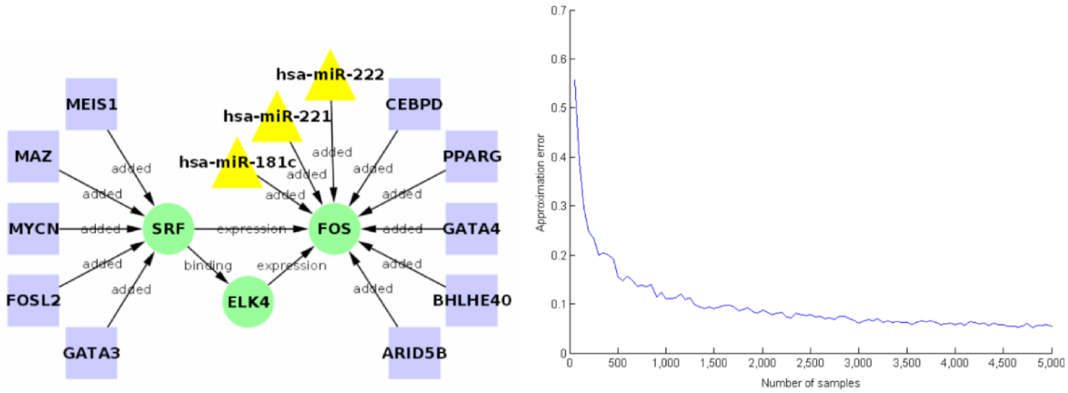
$$P(X_1, X_2, ..., X_{n+s}) = \prod_{i=1}^{n+s} P(X_i|parents(X_i)). \tag{3.2}$$

Because the size of the CPT for each variable X_i is exponential on the number of parents of X_i, this computation is prohibitive for large networks. To complicate matters further, we want an answer to the question: what is the probability of $X_i = x$ given the evidence $\boldsymbol{e}$? This requires the marginalization of X_i from equation Equation 3.2. Since exact inference is computationally infeasible, we have to find an approximation to the marginal probability $P(X_i|\boldsymbol{e})$. In our work, this was achieved by using a Gibbs sampler. The marginal probabilities for all unobserved nodes were sampled at a rate proportional to Q × the number of nodes in the BN, with Q = 250. The BN creation, Gibbs sampler, inference engine and marginalization of nodes were implemented with the `Bayes Net toolbox` for Matlab (Murphy, 2001).

To summarize, each BN was given two different sets of evidence corresponding to two scenarios. In scenario #1, the evidence was the discrete values in ER+ samples of the differentially expressed genes. After providing the BN with the evidence, we ran the inference process and approximated the marginals for all unobserved nodes. We repeated this process for scenario #2, but this time we used as evidence the discrete values in ER- samples.

3.3.12 Approximation of marginal probability for Bayesian inference

In order to empirically determine the value of Q, i.e., the number of samples to draw while using the Gibbs sampler in the estimation of the marginals, we proceeded to create two toy BNs of 16 and 36 nodes. The 16-node network was based on three nodes from the MAPK signaling pathway. These three nodes were subjected to all the steps in our methodology: pathway preprocessing, prediction of target TFs and microRNAs, RF classification and variable importance and, finally, pathway extension. The three pathway nodes (in green) with the TFs (squares) and microRNAs (triangles) that target them are shown in Figure 15(a).



(a)

(b)

Figure 15. (a) Toy BN of 16 nodes and (b) its error in approximating marginals using Gibbs sampler.

The 16- and 36-node networks were small enough that the full joint probabilities could be computed precisely. Therefore, all marginals were computed in an exact manner. We then approximated the marginals using a Gibbs sampler and the approximation error was determined for different number of iterations of the sampler. For the 16-node BN it can be seen in Figure 15(b) that there is little oscillation of the error, and that after 4,000 samples the error stays below 0.05. Our empirical Q = 4,000 / 16 = 250 was used to determine how many samples had to be taken per node. A similar analysis was done with the 36-node network arriving to a similar value of Q. For the 36-node network we continued testing the number of samples up to 50,000 to show how the approximation error continues to decrease (See Figure 59 in the Appendix).

3.4 Results of the inference process using breast cancer data

We have systematically constructed BNs for all the 34 KEGG pathways based on the procedures described in the previous sections. The numbers of nodes and edges in the original pathways and the number of nodes and edges in the expanded Bayesian networks are provided in the Appendix, Table XXXIII.

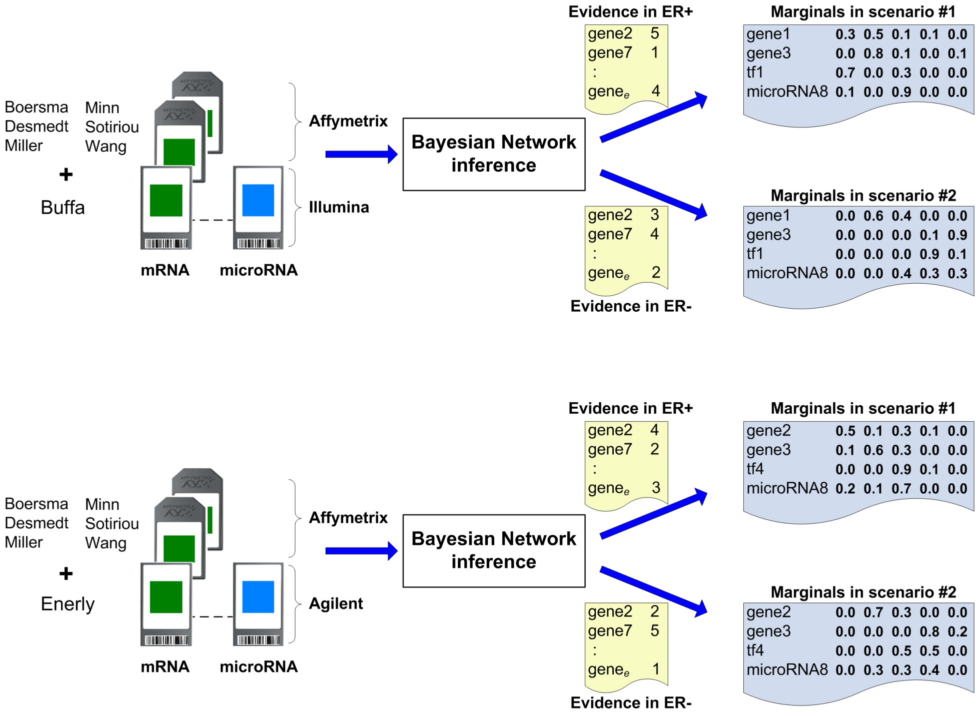
We present our inference results in an attempt at uncovering the relationships among TFs, microRNAs and pathway genes that are associated with ER+ and ER- breast tumors.

As detailed in section 3.3.11, each BN of a pathway was given two different sets of evidence corresponding to two scenarios. In scenario #1, the evidence was the discrete values in ER+ samples of the differentially expressed genes. After providing the BN with the evidence we ran the inference process and approximated the marginals for all unobserved nodes. In scenario

#2, the same inference process was performed and the marginals were approximated. In this case, the evidence used was the discrete values in ER- samples of differentially expressed genes.

In addition to these two scenarios, we created two BNs for each pathway: one BN using the first 6 datasets + Buffa and another using the first 6 datasets + Enerly. Although many nodes in each BN had the same CPTs, those nodes that had a microRNA as parent had their CPTs derived from a different dataset (either Buffa or Enerly respectively).

Our goal in creating these two BNs was to provide further validation to our predictions. If our inference process reports a TF or microRNA as a highly probable regulator, and this result coincides in both the Buffa- and Enerly-derived BNs, it provides greater confirmation that our prediction is plausible. Figure 16 depicts the flowchart of the analysis to create two BNs on which to run inference using the ER+ and ER- scenarios.



Evidence in ER+

gene2	5
gene7	1
:	
gene$_e$	4

Marginals in scenario #1

gene1	0.3	0.5	0.1	0.1	0.0
gene3	0.0	0.8	0.1	0.0	0.1
tf1	0.7	0.0	0.3	0.0	0.0
microRNA8	0.1	0.0	0.9	0.0	0.0

Marginals in scenario #2

gene1	0.0	0.6	0.4	0.0	0.0
gene3	0.0	0.0	0.0	0.1	0.9
tf1	0.0	0.0	0.0	0.9	0.1
microRNA8	0.0	0.0	0.4	0.3	0.3

gene2	3
gene7	4
:	
gene$_e$	2

Evidence in ER-

Evidence in ER+

gene2	4
gene7	2
:	
gene$_e$	3

Marginals in scenario #1

gene2	0.5	0.1	0.3	0.1	0.0
gene3	0.1	0.6	0.3	0.0	0.0
tf4	0.0	0.0	0.9	0.1	0.0
microRNA8	0.2	0.1	0.7	0.0	0.0

Marginals in scenario #2

gene2	0.0	0.7	0.3	0.0	0.0
gene3	0.0	0.0	0.0	0.8	0.2
tf4	0.0	0.0	0.5	0.5	0.0
microRNA8	0.0	0.3	0.3	0.4	0.0

gene2	2
gene7	5
:	
gene$_e$	1

Evidence in ER-

Figure 16. Inference process to generate results.

When analyzing the marginals we obtained, we decided to focus on nodes that fulfilled any of the following two conditions:

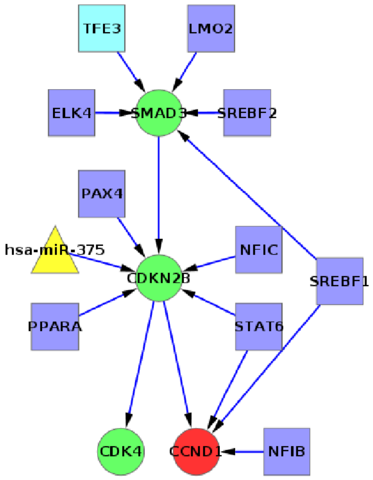
- the node's marginals had one state with a probability larger than 0.8 in scenario #1 and lower than 0.8 in scenario #2 (or vice versa).

- at least one of the nodes marginals for one state had a 2-fold variation in probability between scenario #1 and scenario #2, with the resulting probability being larger than 0.5.

There is no particular reason why we chose these threshold values. They are in fact very stringent and served the purpose of providing a reduced set of results that were easy to manually validate against the true KEGG pathway structure.

3.4.1 Results for the Cell Cycle pathway

In the cell cycle pathway (KEGG Id 04110) we had 9 differentially expressed genes that were obtained from our differential expression analysis. One of those genes, CCND1 (Cyclin D1), was over-expressed in ER+ samples. Being over-expressed in ER+ means that the expression level of CCND1 in ER+ samples was larger than that in ER- samples, in a statistically significant way. Figure 17(left) shows a subset of nodes in the BN created for the cell cycle pathway. In the figure, CCND1 is marked in red to indicate that it is differentially expressed. Figure 17(right) shows how the discrete value of CCND1 in scenario #1, when the discrete value corresponding to ER+ is used as evidence, is larger than the discrete value in scenario #2, when the discrete value corresponding to ER- is used instead. It goes from *very low* in #2 to *medium* in #1. Figure 17(right) also shows the marginals for the rest of the nodes in Figure 17(left) when the 6 datasets + Buffa were used to create the BN (Table XXXIV in the Appendix shows the marginals for the 6 datasets + Enerly). These marginals, for each scenario, indicate the most probable state in which the expression of a gene, TF or microRNA might be, based on the evidence entered in that scenario.



Node	Marginals									
	Scenario #1					Scenario #2				
	very low	medium low	medium	medium high	very high	very low	medium low	medium	medium high	very high
SMAD3	**0.29**	**0.39**	0.18	0.09	0.06	0.21	0.31	0.19	0.12	0.16
CDKN2B	**0.99**		0.01		0.01	0.48	0.11	0.12	0.14	0.14
CCND1			**(de)**			**(de)**				
CDK4	0.01		0.01	0.82	0.16	0.12	0.2	0.18	0.28	0.22
TFE3			**0.69**	0.31			0.04	0.31	0.65	0.01
LMO2		0.06	0.82	0.12	0.01	0.01	0.14	0.77	0.08	
ELK4	0.98	0.01		0.01		0.98		0.01		0.01
SREBF2	0.01		0.99			0.09	0.01	0.89	0.01	
PAX4	1.0					1.0				
NFIC	0.22	0.36	0.35	0.07		0.14	0.36	0.29	0.19	0.01
STAT6	0.01	0.01	0.09	0.9				0.09	0.9	0.01
SREBF1				0.98	0.01			0.07	0.9	0.03
NFIB	0.12	0.26	0.32	0.23	0.08	0.13	0.29	0.29	0.21	0.07
PPARA	0.99	0.01				0.98	0.01		0.01	0.01
hsa-mir-375	0.04	0.1	0.11	0.33	0.42	0.04	0.07	0.15	0.33	0.42

Figure 17. (left) Selected nodes from the merged cell cycle pathway. The original nodes in the pathway are in green. The TFs (squares) and microRNAs (triangles) that target them are also included. The differentially expressed gene CCND1 is marked in red and the TF TFE3 (putative regulator) is in light blue. According to the pathway definition in KEGG, SMAD3 promotes the expression of CDKN2B and CDKN2B inhibits CCND1 and CDK4. According to our analysis, CCND1 was over-expressed in ER+ samples. (right) **de**: the gene is differentially expressed and was used as evidence. The dotted line separates the nodes between those reported using our selection criteria (top part of the table) and others included only to illustrate that their marginals did not change much between scenarios.

When inspecting the TFs: NFIB, STAT6 and SREBF1 that from sequence analysis and RF we have predicted to target CCND1 directly, we realize that their marginals are very similar in both scenarios. Because we know that the expression of CCND1 changed between scenarios #2 and #1, we are looking for a TF or microRNA that may also have changed between those scenarios and that may help explain the change in expression for CCND1. Neither of the TFs or microRNAs that target CCND1 have a significant change in their marginals between scenarios and this is why they are not depicted in Figure 17(left).

The TF TFE3 (transcription factor binding to IGHM enhancer 3) may provide a better explanation of why CCND1 is differentially expressed, even if TFE3 does not target CCND1 directly. In Figure 17(left) TFE3 is in light blue and targets SMAD3. Between scenarios #2 and #1 we can see – Figure 17(right)– that there is more certainty in scenario #1 that SMAD3 is at a lower state (a combined *very low* and *medium low* of 0.29+0.39=0.68). This implies a lower level of expression in that scenario (vs. 0.21+0.31=0.52 in scenario #2). The marginals have a moderate change from higher expression states in scenario #2 to lower states in #1. This transition is much sharper for Enerly (See Table XXXIV in the Appendix). In the Cell cycle pathway, SMAD3 promotes the expression of CDKN2B, which in turn regulates the expression of CCND1 and CDK4 by inhibiting them. Our BN simply keeps directed edges between nodes but is not aware of the semantics of each edge (inhibition, expression, and so forth). Nevertheless our results adjust very well to the semantics of the pathway. When SMAD3 switches to a lower state (from scenario #2 to #1), CDNK2B has also a sharp decrease of expression to a *very low* state (with an increase of certainty from 0.48 in scenario #2 to 0.99 in scenario #1). Therefore, with a high chance of having low expression of CDKN2B, we also have a high chance of not inhibiting neither CCND1 nor CDK4 and this results in an increase in their expression levels (for CCND1, from *very low* in scenario #2 to *medium* in #1; and for CDK4 it goes from a somewhat uncertain state of expression in scenario #2 to a 0.82 certainty of having *medium high* expression level in scenario #1).

Upon reviewing the TFs that are predicted to target SMAD3, we see that TFE3 is the only one with a marked contrast between scenarios. In scenario #2 there is 0.65 probability that its expression is *medium high* but this probability decreases to a 0.31 (more than 2-fold decrease) in scenario #1. This sharp decrease occurs because in scenario #1 there is more certainty of TFE3 being in a *medium* state of expression (0.69 vs. 0.31 in scenario #2). We therefore hypothesize that the transcription factor TFE3 is a key regulator in the Cell cycle pathway when comparing ER- and ER+ samples. We are not implying by any means that TFE3 affects directly the expression of SMAD3 but there is a clear relationship between their changes in expression levels and this allows us to postulate TFE3 as a regulator in the pathway. In fact, TFE3 is a well-known transcription factor (Beckmann et al., 1990) and there is ample evidence of its synergizing effects with SMAD3 to enhance Transformer Growth Factor β (TGF-β) dependent transcription (Hua et al., 1999; Hua et al., 2000).

3.4.2 Results for the p53 signaling pathway

The analysis of the p53 signaling pathway (KEGG Id 04115) provides an example of how to identify a regulator based on direct interactions between the regulator and genes in the pathway. For this pathway, our differential expression analysis reported 8 differentially expressed genes. Very few TFs passed our selection criteria and only one of them overlapped between the Buffa and Enerly datasets. This is the case of STAT5B known as signal transducer and activator of transcription 5B. STAT5B was predicted to target only 2 genes in this pathway: IGFBP3 (insulin-like growth factor binding protein 3) and PERP (p53 apoptosis effector related to PMP-22). These two genes are located in different parts of the pathway and are not directly related to each other.

The marginals corresponding to IGFBP3 do not seem to have much of a variation between our two scenarios (Table XXXV below and Table XI in the Appendix). In contrast, PERP is differentially expressed (down-regulated) in ER+ samples, i.e., scenario #1. We can see that the TF STAT5B shifts its marginal probability of being in a somewhat uncertain state of *medium low* expression in scenario #2 to a more certain state in scenario #1 (*medium low* = 0.84). This shift is accompanied, in scenario #1, by a decrement of the marginal corresponding to the lowest level of expression (*very low*) which can be interpreted as a subtle increase of expression of STAT5B in scenario #1 with respect to scenario #2.

TABLE XI

SELECTED MARGINALS FOR THE P53 SIGNALING PATHWAY (6 DATASETS + ENERLY)

Node	Marginals									
	Scenario #1					Scenario #2				
	very low	medium low	medium	medium high	very high	very low	medium low	medium	medium high	very high
STAT5B	0.16	**0.84**				0.24	0.73	0.02		
PERP			**(de)**						**(de)**	
IGFBP3	0.23	0.23	0.2	0.14	0.21	0.2	0.2	0.23	0.15	0.21

STAT5 is one of the seven members of the STAT (signal transducers and activators of transcription) family of TFs and mediates the responses of cytokines, growth factors and hormones (Basham et al., 2008). It has been shown that STAT5 regulates apoptosis in a wide range of tumor cells (Longley and Johnston, 2007). STAT5A and STAT5B are different proteins encoded by different genes.

PERP, a p53 transcriptional target, is induced specifically during apoptosis but not cell cycle arrest. Down-regulation of PERP is associated with the aggressive, monosomy 3-type of uveal melanoma (UM) (Davies et al., 2011). It has not been proven that PERP is a direct target of STAT5B (Basham et al., 2008). But through our Bayesian inference process we were able to determine that STAT5B (by interacting with PERP) might be a key regulator in the p53 signaling pathway. This result was validated by two BNs constructed with different datasets (Buffa and Enerly).

3.4.3 Results for the MAPK signaling pathway

In addition to the results mentioned in the previous section, here we detail other findings obtained for a different pathway: the MAPK signaling pathway (KEGG Id 04010), which are also related to the TF STAT5B mentioned before.

As it was the case for the p53 signaling pathway, here we also had the chance to identify a regulator based on direct interactions between the regulator and genes in the pathway. We had 15 differentially expressed genes reported for this pathway. Here too, our strict selection criteria yielded few TFs (only 5 in Enerly). Yet, one of them, STAT5B might be considered a regulator for this pathway based on our results.

We predicted 6 target genes of STAT5B: MAP3K12, NFKB2, RRAS2, FGF23, MAPK10 and PTPRR. These genes had no edges connecting them directly in the pathway. The marginals corresponding to the last 3 genes do not vary much between scenarios #1 and #2 (below the dotted line in Table XII and Table XXXVI in the Appendix). In contrast, the first 3 genes do show a difference between the scenarios and MAP3K12 is in fact differentially expressed and up-regulated in ER+ samples. The TF STAT5B shifts its certainty of being in state *medium low* in scenario #2 to a more uncertain state in scenario #1. In scenario #1 we see a slight increment in the marginals corresponding to higher levels of expression with respect to scenario #2. The fact that STAT5B may slightly increase its expression in scenario #1 provides a better explanation of why NFKB2 and RRAS2 also have a shift in marginals towards higher expression states from scenario #2 to #1. Therefore, we postulate that STAT5B is a potential regulator of the MAPK signaling pathway in breast cancer when comparing ER+ and ER- patients. Because STAT5B showed to be of importance in two different pathways, this reinforces our hypothesis that STAT5B is a key regulator in breast cancer patients.

TABLE XII

SELECTED MARGINALS FOR THE MAPK SIGNALING PATHWAY (6 DATASETS + ENERLY)

Node	Marginals									
	Scenario #1					Scenario #2				
	very low	medium low	medium	medium high	very high	very low	medium low	medium	medium high	very high
STAT5B	0.22	0.64	**0.09**	**0.03**	**0.01**	0.1	0.88	0.01	0.01	
MAP3K12			**(de)**			**(de)**				
NFKB2	0.18	**0.22**	**0.22**	**0.2**	**0.18**	0.74	0.09	0.06	0.06	0.05
RRAS2	0.4	**0.14**	**0.16**	**0.14**	**0.16**	0.78	0.07	0.06	0.06	0.04
PTPRR	0.27	0.17	0.2	0.21	0.15	0.17	0.18	0.18	0.23	0.24
FGF23	0.15	0.24	0.18	0.2	0.23	0.19	0.2	0.19	0.2	0.21
MAPK10	0.17	0.22	0.23	0.22	0.16	0.18	0.15	0.24	0.2	0.23

3.5 Conclusions

We proposed an integrative bioinformatics methodology that combines: a) the TFs and microRNAs that are predicted to target pathway genes, with b) microarray expression profiles of mRNA and microRNA, in conjunction with c) the known structure of molecular pathways. All these elements were integrated into a probabilistic framework (BN) that was used to make inferences about key TFs and microRNAs as regulators of the pathway. Using the procedures described in this chapter, one can systematically construct a BN for each individual pathway of interest. We have utilized 8 microarray expression datasets of mRNA and microRNA on ER+ and ER- breast tumors to demonstrate how to use the differentially expressed genes as evidence in order to infer key regulators in the constructed BNs.

Another important use of our framework is to propose hypotheses about the expression levels of TFs or microRNAs and their effect on genes. We foresee the researcher posing questions of the form: "What would the expression level of genes g_1 and g_2 be if $microRNA_3$ is expressed at a very high level?"

Several technical issues deserve further investigation. When making inference about the expression level of a gene, TF or microRNA, we would ideally want to obtain the most probable explanation (MPE) given the evidence at hand. This evidence can be tangible, i.e., obtained from a microarray experiment, or, as it was mentioned before, it can be a set of hypotheses that interest us. In either case, an exact solution to the MPE problem in Bayesian inference has proven to be elusive due to the fact that approximating the MPE or finding the k-th MPE are both NP-hard problems (Jensen and Nielsen, 2007). Thus, in this work we have decided to use the marginals as a proxy for MPE. In turn, we approximated the marginals for the unobserved nodes using a stochastic sampling algorithm (Gibbs sampler). A possible way to improve our methodology will be to thoroughly examine different importance sampling algorithms that will minimize the variance between the drawn samples and the target distribution (Cheng and Druzdzel, 2000).

Finally, a self-imposed limitation of our model was the removal of edges that would create cycles in the network. Another possible step towards improving our probabilistic framework is to use a dynamic Bayesian network (DBN), which allows for cycles, and that better reflects the positive feedback present in many molecular pathways.

CHAPTER 4

A COMPUTATION PIPELINE FOR THE INTEGRATIVE ANALYSIS OF METHYLATION AND MRNA DATA.

4.1 Introduction

Cytosine-guanine dinucleotides (CpGs=5'-CG-3') throughout our genome are chemically tagged with methyl groups, which serve to regulate gene expression both globally and locally (Suzuki and Bird, 2008). CpG sites do not occur as frequently across the genome of vertebrates as expected from the proportion of C and G base pairs. CpG sites are only present at $\frac{1}{5}$ of their expected number although in certain genomic regions their density reaches up to 10 times of what is observed in the rest of the genome (Lewin et al., 2011). These regions are called **CpG islands** and there are approximately 29,000 of them in the human genome. Transcription of genes can be affected by methylation of CpG sites in two different ways:

- a methylated CpG site that corresponds with the binding site of a transcription factor (TF) may prevent binding of the TF.
- methylated CpG sites may facilitate the binding of repressors.

The rapid evolution of techniques to study this epigenetic mark now allow us to analyze genome-wide CpG methylation sites, but we are pressed by the need to discern the biological information conveyed from not only their magnitude of change, but also their positional context within the genome. Growing evidence suggests that CpG islands escape methylation, while more isolated CpGs within the genome are more variably methylated (Greally, 2013). Additionally, genomic loci showing cell- and tissue-specific patterns of differential methylation are more often found in regions with reduced CpG density, supporting the notion that methylation of CpG sites outside of the classical CpG islands play an important role in the regulation of gene expression (Pollard et al., 2009).

4.2 Preliminaries

It becomes imperative then to find analysis pipelines with the ability to seemlessly integrate methylation and mRNA data. However, existing bioinformatics tools, such as the R packages `lumi` (Du et al., 2008), `IMA` (Wang et al., 2012) and `SWAN` (Maksimovic et al., 2012), focus on the differential methylation analysis of individual CpGs in the Illumina 27K/450K high resolution methylation arrays. COHCAP (Warden et al., 2013), a more recently released pipeline for the integrative analysis of methylation and gene expression data, addresses the possible involvement of multiple CpGs in gene regulation by taking the average methylation levels of differentially methylated CpG probes in the proximity of a CpG island. Due to aggregation, the positional effect on gene regulation of individually methylated CpGs may become obscure in this approach. Here we present a workflow named `me-mRNA-pipe` to:

- Comprehensively examine the relationship between phenotypes and positional methylation levels of CpGs
- Further evaluate the potential association of changes in gene expression with a combination of CpG methylation levels.

We address these two main questions in order to provide results that are considered more biologically relevant. Our tool is applicable to the analysis of several widely used methylation arrays in addition to various types of genome-wide gene expression arrays.

The majority of the work that focuses on finding a linkage between methylation and gene expression uses regression methods, machine learning strategies or a combination of both. Cheng and Gerstein (Cheng and Gerstein, 2011) use support vector regression to identify how methylation of histones in the proximity of transcription start sites and transcription end sites affect the expression of genes. Their regression model attempts to predict the expression of genes based on the amount of methylation in their sites of interest. Another approach, more closely related to the one we propose, is the one described by Sun and Wang (Sun and Wang, 2012). The authors analyze DNA methylation data obtained from the promoter regions of genes and correlate it with the expression levels of the genes. They apply penalized logistic regression concurrently on all genes and their methylated promoters.

As the technology to measure methylation improves and researchers have access to more sophisticated experimental tools, there is a pressing need for bioinformatics methods that can help identify what CpG sites are most responsible in affecting the expression of certain genes.

4.3 General features

The analysis pipeline we developed can process three possible types of expression data:

- Raw microarray data files
- Standard Matrix-series files downloaded from the Gene Expression Omnibus (GEO) (Edgar et al., 2002)

- Data exported from third-party tools.

Currently, `me-mRNA-pipe` can only process a limited number of mRNA and methylation array platforms from raw data but a much wider variety of microarrays can be imported from GEO or other external tools. The software provides the flexibility to process user-defined microarray platforms as long as they are accompanied by an appropriate platform definition file. These and other detailed specifications can be found in a user manual that was created for the pipeline (not included in this document). Once the data have been imported into the pipeline, its overall execution can be divided into four major steps:

- *Descriptive statistics and plots*: At this stage of the pipeline, methylation and mRNA data are processed separately. The microarray samples are displayed in a) 2-dimensional plots using the first two components of a principal component analysis (PCA) and b) dendrograms with hierarchical clustering. These give the user a quality control measure in terms of how well the samples cluster between the phenotypes. Additionally, density plots are provided for all samples to assess the effect of normalization.

  In the case of methylation, the *beta* values are used for these analyses whereas, for mRNA, the expression values from the probes are considered. Lists of differentially expressed probes and differentially methylated CpGs can be input from an external source, or can be computed by the pipeline.

- *Associating differentially expressed genes and differentially methylated CpGs*: Using platform definition files, the differentially expressed mRNA probes are converted to differentially expressed transcripts with unique RefSeq Ids. If more than one transcript for the same gene is differentially expressed, the transcript with the smallest adjusted p-value is selected. The genomic coordinates of these transcripts are then used to determine the overlap with differentially methylated CpGs. The overlap between CpGs and genes lays out the basis for the integrative analysis conducted by the pipeline. Based on their genomic coordinates, CpGs are grouped by location. A location of a CpG can be with respect to the gene with which it overlaps, or with respect to a known CpG island, or a user-defined location. These locations are normally obtained from the microarray manufacturer's manifest.

- *Interactions of CpGs with phenotypes and other positional statistics*: The first part of the integrative analysis determines interactions between differentially methylated CpGs and the phenotypes of the differentially expressed genes. We conduct, per gene, a two-factor ANOVA analysis of the beta values of the CpGs associated with the gene. One factor of the analysis is the phenotype and the other is the CpG location. This allows us to rank differentially expressed genes based on the p-values of the interactions from their overlapping CpGs (Supplementary Methods).

  Another view at how CpGs relate to expression values is obtained from the correlation analysis implemented in the pipeline. In this analysis, Spearman correlation coefficients are computed between the expression values of a gene and the beta values of its CpGs. After the coefficients have been calculated for all CpGs and genes, Fisher's exact tests are conducted to determine if the proportion of positively and negatively correlated coefficients at specific locations is the same as the global proportion (Supplementary Methods).

- *Modeling gene expression with CpG beta values:* For a gene with multiple overlapping CpGs, we model the expression value of the gene with the beta values of the individual CpGs using multiple linear regression. Additionally, when the number of CpGs is large we use a LASSO penalized regression model (Tibshirani, 2011) which aims at determining a subset of CpGs that are "the most important" in explaining the expression levels of the gene.

4.3.1 Methylation microarrays

The pipeline benefits from the whole-genome coverage of Illumina's and NimbleGen's high resolution methylation arrays:

- HumanMethylation450 BeadChip array (450K): genome-wide coverage of different gene regions and CpG islands. It covers 99% of RefSeq genes.
- HumanMethylation27 BeadChip array (27K): targets CpG sites located in the proximal promoter regions of more than 14,000 genes and more than 100 microRNAs.
- NimbleGen 385K Human RefSeq promoter array: covers the distal and proximal promoters of 100% RefSeq genes. Probes in the promoter region are separated by 100 bp.

Details about the coverage of the Illumina arrays can be found in Table XIII (Illumina, 2012) and Table XIV (Illumina, 2010). Other platforms can be supported by creating the appropriate platform definition files.

TABLE XIII

COVERAGE OF THE ILLUMINA HUMANMETHYLATION450 BEADCHIP ARRAY

	Location	**Genes mapped**	**Percentage of genes covered**
Coding protein	TSS1500	17,820	94%
	TSS200	14,895	79%
	5'UTR	13,865	78%
	1stExon	15,127	80%
	Body	17,071	97%
	3'UTR	13,042	72%
Non-coding RNA (microRNA)	TSS1500	2,672	88%
	TSS200	1,967	65%
	Body	2,345	77%
	Location	**Islands mapped**	**Percentage of islands covered**
CpG Island	North Shelf	23,896	86%
	North Shore	25,770	93%
	Island	26,153	94%
	South Shore	25,614	92%
	South Shelf	23,968	86%

4.3.2 Illumina 450K methylation array

This array has 482,421 cytosine probes throughout the entire human genome. Each CpG can have three different location contexts: a) location with respect to a gene, b) location with respect to a CpG island and c) custom (user-defined) location.

Location by gene Although some probes are located in gene desert areas, the probes that overlap with a known gene have a similar layout to the one depicted in Figure 18.

The CpG probe locations with respect to a gene have been categorized by the manufacturer as: TSS1500, TSS200, 5'UTR, 1stExon, Body and 3'UTR, where:

TABLE XIV

COVERAGE OF THE ILLUMINA HUMANMETHYLATION27 BEADCHIP ARRAY

	Genes mapped	**Average coverage** (in number of sites)
RefSeq genes	14,475	1.9
Hot spots in cancer genes	144	7.6
Cancer-related targets	982	1.9
microRNA promoters	110	2.3

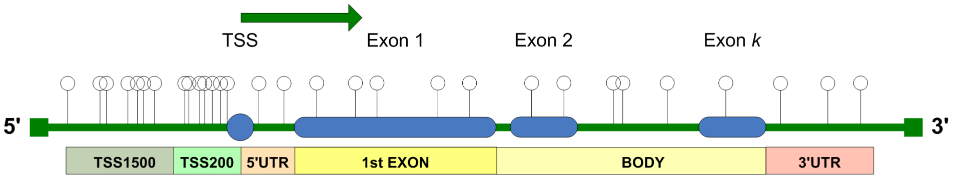


Figure 18. Location by gene: CpG probes –marked as circles with sticks– that overlap with a gene

- TSS is the transcription start site of a gene and the numbers 1500 and 200 refer to the maximum number of base pairs from the TSS.
- UTR is an untranslated region at the 5' end or 3' end of the coding region of a gene.
- 1stExon refers to the first exon of a gene.
- Body is the remainder, including the introns and exons after the first exon and up to the start of the 3'UTR.

Location by CpG island In addition to (possibly) overlapping with the promoter of a gene or with its coding region, a CpG probe will always have a location with respect to a known CpG island. This allows the same CpG probe, for example, to be in the promoter of a gene (say, gene_location = TSS200) and within the boundaries of a CpG island (say, cpg_location = Island). It is this richness of location information that `me-mRNA-pipe` exploits. Figure 19 shows the different locations of a CpG with respect to an island.

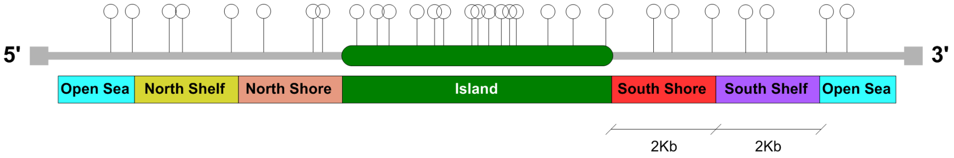


Figure 19. Location by CpG island: CpG probes overlapping or in the vicinity of an island.

The locations with respect to a CpG island provided by the manufacturer are: N_Shelf, N_Shore, Island, S_Shore, S_Shelf and Open_sea.

- N_Shore and S_Shore are the neighboring regions of the CpG island for up to 2 Kb. The prefixes “North” or “South” refer to the 5’ end and 3’ end of the chromosome respectively.
- The Shelf (also “North” and “South”) is the neighboring region of the shore that extends up to 4 Kb away from the CpG island.

- Open_sea is used when the CpG is more than 4 Kb away from a CpG island.

Custom location The previous two locations are implemented in the pipeline as defined by the manufacturer. A third type of location can be customized by the user and, by default in this platform, has been set to a more compact version of the *location by CpG island*. Previously, the shores and shelves were characterized as "North" or "South". But this distinction between "North" and "South" referred to the 5' end and 3' end of the chromosome respectively. The distinction loses its true meaning if we wish to integrate a *location by CpG island* with a *location by gene* because genes have an orientation and are either located in the + or - strand. Therefore the third type of location we devised combines shore and shelves (distance of 4 Kb from an island) and does not distinguish between north and south. Figure 20 illustrates this type of location.

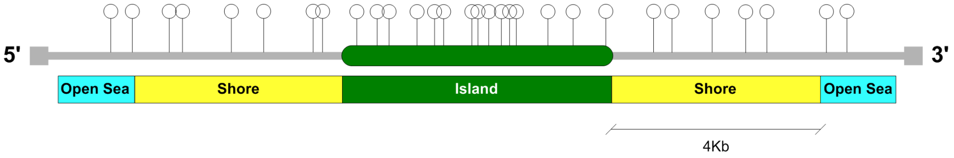


Figure 20. Custom location: Location relative to a CpG island with combined shores/shelves and no distinction about orientation.

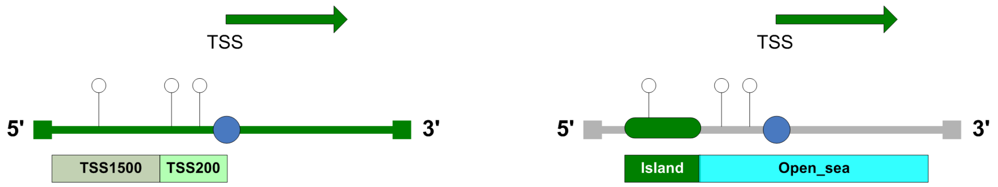
4.3.3 Illumina 27K methylation array

It contains 27,578 CpG probes present in the vicinity of the TSSs of genes and 90% of these probes are also included in the 450K array described before. The types of location for this platform are described below and illustrated in Figure 21.

Location by gene Possible locations are: TSS200 for CpGs located up to 200 bp upstream of a TSS; or TSS1500 if the distance is greater than 200 bp and less than 1,500 bp upstream from the TSS.

Location by CpG island The locations are: Island if the CpG overlaps with a CpG island; or Open_sea if it does not overlap with a known island.

Custom location Provides more granularity than the locations by gene. Possible locations are TSS200, TSS600, TSS1000 and TSS1500 depending on their distance to the TSS.



(a) Location by gene.

(b) Location by CpG island.

Figure 21. Location of CpGs in the Illumina 27K array. As it was the case in the 450K platform, the same CpG can have a location by gene and by CpG island.

4.3.4 NimbleGen 385K Human RefSeq promoter array

This array has 711,794 cytosine probes throughout the promoters of all known RefSeq genes. The three types of locations a CpG can have refer to the TSS of the closest gene.

Location by gene #1 Locations are divided into upstream and downstream from the TSS and have the prefix "UP_" and "DOWN_" respectively. The suffix indicates the distance to the TSS. For example, UP_TSS200 and DOWN_TSS200 refer to CpGs located 200 bp upstream/downstream of a TSS respectively; UP_TSS1Kb and DOWN_TSS1Kb if the distance is greater than 200 bp and less than or equal to 1,000 bp upstream/downstream; and similarly for the rest. Figure 22 shows all these locations.

Location by gene #2 Uses the same distances as the in the previous case but no distinction is made if the CpG is upstream or downstream from the TSS. The locations are: TSS200, TSS1Kb, TSS2Kb and TSS10Kb.

Location by gene #3 Provides a more compact view of the previous case, dividing the CpGs in proximal or distal to the TSS. Possible locations are TSS2Kb and TSS10Kb.

Analysis results based on the Illumina 27K and 450K methylation arrays have revealed that a differentially expressed gene can be associated with multiple differentially methylated CpG probes. Among these probes, some may be hyper-methylated and others may be hypomethylated, a phenomenon that can be explained by the existence of statistical interactions between the phenotypes and the locations of these CpGs. One of the objectives of our pipeline is the detection of such interactions. Another objective is to find, for a given gene, a subset of

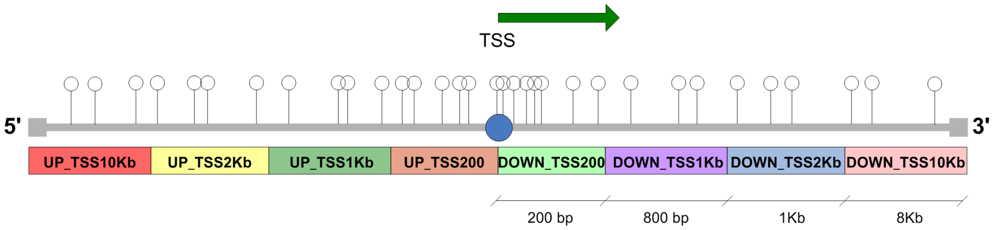


Figure 22. Location by gene #1: CpG probes –marked as circles with sticks– that overlap with a gene

differentially methylated CpGs that can best explain the observed expression level of the gene. The identified CpGs can serve as unique epigenetic fingerprints related to gene expression.

Our pipeline consists of four major modules: i) Data pre-preprocessing, ii) Generation of descriptive statistics and plots; iii) Detection of interactions of CpGs with phenotypes and other positional statistics; and iv) Modeling of gene expression using CpG beta values. A flowchart of the pipeline is shown in Figure 23.

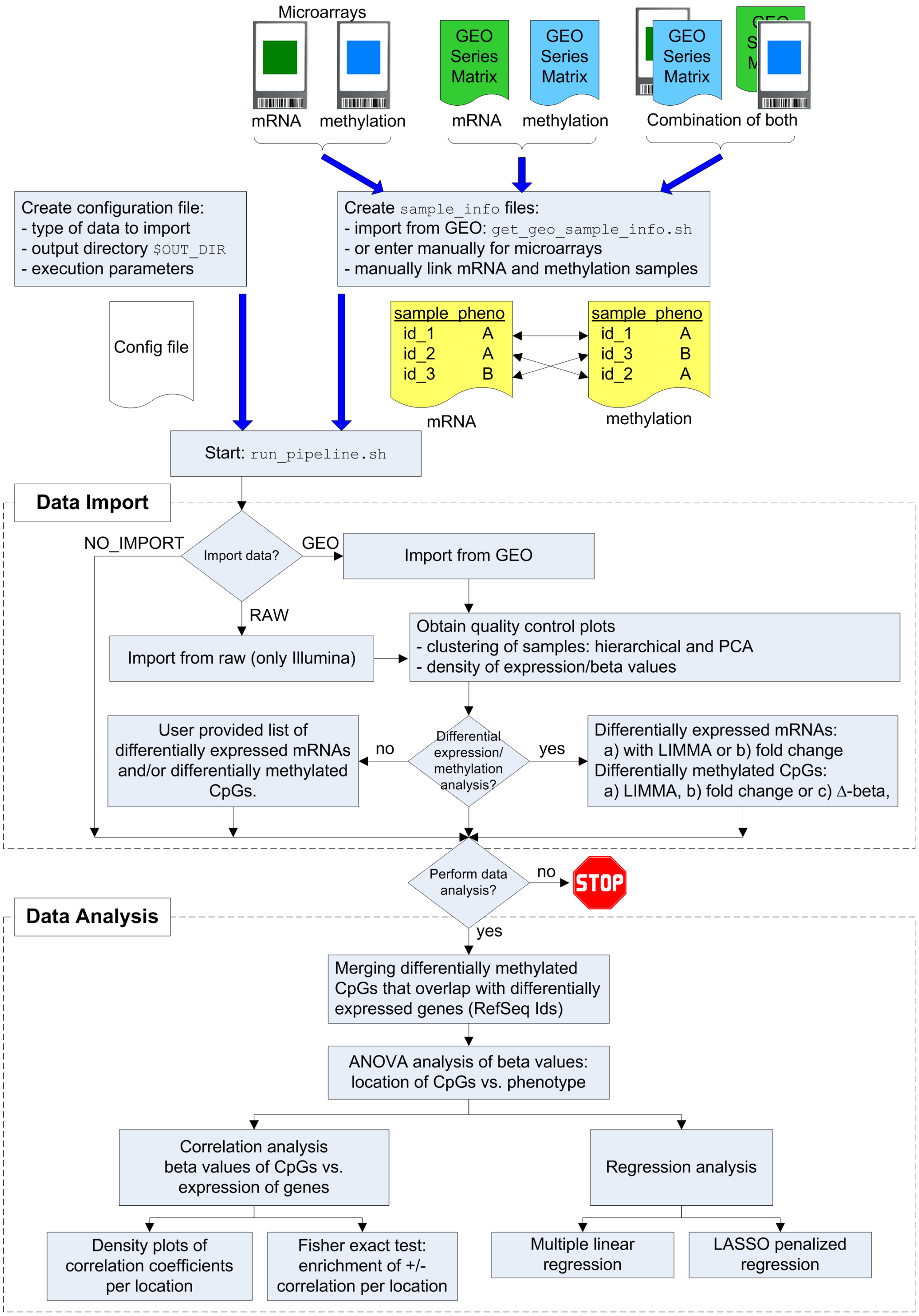

Microarrays
mRNA
methylation
GEO Series Matrix
GEO Series Matrix
mRNA
methylation
GEO Series Matrix
Combination of both
Create configuration file:
- type of data to import
- output directory $OUT_DIR
- execution parameters
Create sample_info files:
- import from GEO: get_geo_sample_info.sh
- or enter manually for microarrays
- manually link mRNA and methylation samples
Config file
sample pheno
id_1 A
id_2 A
id_3 B
mRNA
sample pheno
id_1 A
id_3 B
id_2 A
methylation
Start: run_pipeline.sh
Data Import
NO_IMPORT
Import data?
GEO
Import from GEO
RAW
Import from raw (only Illumina)
Obtain quality control plots
- clustering of samples: hierarchical and PCA
- density of expression/beta values
User provided list of differentially expressed mRNAs and/or differentially methylated CpGs.
no
Differential expression/ methylation analysis?
yes
Differentially expressed mRNAs:
a) with LIMMA or b) fold change
Differentially methylated CpGs:
a) LIMMA, b) fold change or c) Δ-beta,
Perform data analysis?
no
STOP
Data Analysis
yes
Merging differentially methylated CpGs that overlap with differentially expressed genes (RefSeq Ids)
ANOVA analysis of beta values: location of CpGs vs. phenotype
Correlation analysis beta values of CpGs vs. expression of genes
Regression analysis
Density plots of correlation coefficients per location
Fisher exact test: enrichment of +/- correlation per location
Multiple linear regression
LASSO penalized regression

Figure 23. Execution flowchart of me-mRNA-pipe

4.4 Proposed method

This section provides the details about the methods implemented in `me-mRNA-pipe`.

4.4.1 Notation

The following notational conventions are adopted for the rest of this chapter:

- β : indicates the methylation level of a CpG. It is called *beta value* and is described in the next section.
- ρ : represents a correlation coefficient, either Spearman's or Pearson's.
- Other Greek letters stand for values/parameters that will be computed or estimated by `me-mRNA-pipe`.
- In an equation with a superscript k in the LHS, k refers to a particular gene. In the same equation, all references to k in the RHS will be omitted as it is clear from the context that the arguments refer to gene k. For example:

  $$\varphi^k = \psi + \omega$$

  where φ^k is a value associated to gene k and our goal is to determine ψ and ω. The equation is per gene, therefore ψ and ω are specific to gene k and the superscript k is omitted.

4.4.2 Beta values to measure methylation levels

`me-mRNA-pipe` measures methylation of a CpG using beta values. Equation 4.1 shows how a beta value is computed at the ith CpG:

$$\beta_i = \frac{\max(M_i, 0)}{\max(M_i, 0) + \max(U_i, 0) + \alpha}, \tag{4.1}$$

where M_i and U_i refer to the signal intensity of the methylated and unmethylated probes assayed for CpG$_i$. It is important to note that, in its original definition by Illumina (Bibikova et al., 2006), Equation 4.1 was $\beta_i = \frac{\max(M_i,0)}{|M_i|+|U_i|+\alpha}$ where $\alpha = 100$ and the absolute values in the denominator corrected for negative signal after background correction.

4.4.3 Data pre-processing

The data import module of the pipeline allows the user to conduct differential expression and/or differential methylation analyses. When data are imported from GEO, the user may run the pipeline with the lists of differentially expressed genes and differentially methylated CpGs that the authors of the original study obtained. Knowing that this is not always possible, the user can opt to conduct differential analysis with `me-mRNA-pipe`.

4.4.3.1 Differential expression of mRNA probes

There are two methods to determine differential expression of mRNAs:

- Fold change: For each mRNA probe, a ratio of the mean expression values in both phenotypes is obtained. If the data were flagged as log2 transformed in the configuration file, then a log2-fold change is computed. Probes with a fold change greater than a threshold (also set in the configuration file) are considered to be differentially expressed. This is only recommended when the sample size is very small.

- Moderated t-statistic controlling for false discovery rate (FDR): The `limma` package in R/`Bioconductor` (Smyth, 2004) is used to obtain, for each probe, an adjusted p-value of the difference in expression between phenotypes. The p-values are adjusted for multiple hypothesis testing to control the FDR with the Benjamini-Hochberg algorithm (Benjamini and Hochberg, 1995). Probes with an adjusted p-value less than a threshold set in the configuration file are reported.

4.4.3.2 Differential methylation of CpG probes

There are three methods to determine differential methylation of CpGs:

- Δbeta: The mean beta values of each CpG probe in both phenotypes are compared against each other. If the absolute difference is greater than a threshold, the CpG is considered to be differentially methylated.

- Fold change: For each CpG, a ratio of the mean beta values in the phenotypes is obtained. If the ratio (fold change) is greater than a threshold set in the configuration file, the CpG is reported.

- Moderated t-statistic controlling for false discovery rate (FDR): The same approach previously described for mRNAs but applied in this case to the beta values of CpGs.

- Combination of the filters: Any combination of the previous conditions has to be met in order to report the CpG as differentially methylated.

4.4.4 ANOVA analysis

The ANOVA analysis is conducted on the beta values of differentially methylated CpGs associated to a differentially expressed gene. The two-factor model is described in Equation 4.2, where β^k_{pq} is the beta value of a CpG associated to gene k; μ is an unknown constant; ϕ_p is the phenotype effect for phenotype $p = 1, 2$ –only two phenotypes are considered; τ_q indicates the location effect for location $q = 1, 2, \ldots, Q$ when considering Q possible locations of a CpG overlapping with a gene/island; $(\phi\tau)_{pq}$ is the interaction effect between the phenotype p and location q; and finally, ϵ_{pq} models other sources of variation in beta values that arise neither from the phenotype nor from the location of the CpG.

$$\beta^k_{pq} = \mu + \phi_p + \tau_q + (\phi\tau)_{pq} + \epsilon_{pq}, \tag{4.2}$$

For each gene k, its CpGs are tested for the following null hypotheses:

1. $H^{\tau}_0 =$ the variation in beta values does not depend on the location of the CpG, i.e., $\tau_q = 0$ for $q = 1, 2, \ldots, Q$
2. $H^{(\phi\tau)}_0 =$ there is no interaction between CpG locations and phenotypes, i.e., $(\phi\tau)_{pq} = 0$ for all possible combinations of $p = 1, 2$ and $q = 1, 2, \ldots, Q$

The alternative hypothesis in 1. is that some $\tau_q \neq 0$, which implies that the variation in beta values is also a product of the CpG's genomic location. Similarly, the alternative hypothesis in 2. is $\exists\ p, q \mid (\phi\tau)_{pq} \neq 0$, which means there is at least one combination of phenotype and location that accounts for the variability of beta values.

It is important to note that we are not interested in testing $\phi_p = 0$ for $p = 1, 2$ because we already know that the CpGs are differentially methylated between the phenotypes.

The pipeline provides the p-values for all these hypotheses tests. All genes are then ranked based on the p-value obtained from the hypothesis test about the interaction $(\phi\tau)_{pq}$ of phenotype and CpG location.

4.4.5 Correlation analysis

The previous analysis focuses solely on the variation of beta values between the phenotypes. From now on we turn our attention at how CpGs relate to gene expression values. The pipeline conducts a correlation analysis in which a Pearson's correlation coefficient ρ^k_{Y,B_j} is computed between the expression values of gene k and the beta values of an overlapping CpG j. The vector Y is defined as $Y = [y_1, y_2, \ldots, y_N]$ and contains the expression values of gene k across all N samples. Similarly, $B_j = [\beta_{1,j}, \beta_{2,j}, \ldots, \beta_{N,j}]$ are the beta values of CpG j in all samples. A value of $\rho^k_j > 0$ indicates a positive correlation between gene k and CpG j, i.e., if the gene is over-expressed in one phenotype, the CpG is hyper-methylated in that phenotype as well. The converse is also true: under-expression in one phenotype corresponds to hypo-methylation in the same phenotype. Alternatively, when $\rho^k_j < 0$ we have over-expression and hypo-methylation or under-expression and hyper-methylation relative to the phenotypes.

After the correlation coefficients are calculated for all CpGs and genes, we can determine if the proportion of positively and negatively correlated coefficients at a specific location is the same as the global proportion. The main question we want to address is: "Are there more positively (or negatively) correlated CpGs at location *loc* than in the rest of the locations?".

In order to answer this generic question, the pipeline conducts a Fisher's exact test at each location to determine if the proportion of positively and negatively correlated coefficients at location = *loc* is the same as the proportion for all other locations ≠ *loc*. To run the test a contingency matrix is created for each location *loc* as shown in Figure 24.

	Number of CpGs	
	with negative correlation ($\rho < 0$)	with positive correlation ($\rho > 0$)
Location = *loc*	n_1	n_2
Other locations	n_3	n_4

Figure 24. Contingency matrix used for Fisher's exact test.

The three types of locations considered in this analysis are the ones described in sections 4.3.2 , 4.3.3 and 4.3.4. For each location type there are different location values, e.g.: if we consider location by gene in the 450K platform, we have $loc \in$ {TSS1500, TSS200, 5'UTR, 1stExon, Body, 3'UTR}.

Correlation density plots: In addition to the previous analysis, a distribution of correlation coefficients is obtained for each location *loc*. For all the CpGs that fall in location *loc*, their correlation coefficients are binned, from -1.0 to 1.0, with bins of size 0.01. A normalized frequency is obtained at each bin by dividing the count of CpGs in the bin by the total number of CpGs. Finally, a smoothing spline is fit to the normalized frequencies thus obtaining a density curve. An example of this plot can be found in Figure 30, section 4.5.3.

4.4.6 Multiple linear regression

To establish the effect of how CpG methylation influences gene expression, a multiple regression analysis is implemented in the pipeline. This model attempts to find a linear relationship between the expression values of a gene and the beta values of the CpGs that overlap with the gene. Equation 4.3 models this relationship.

$$y_i^k = \gamma_0 + \sum_{j=1}^{m} \beta_{ij}\gamma_j + \epsilon_i, \text{ with } i = 1, 2, \ldots, N, \tag{4.3}$$

where $\boldsymbol{\gamma}^k = (\gamma_0, \gamma_1, ..., \gamma_m)$ are the regression coefficients; N is the number of samples; m is the number of differentially methylated probes associated to gene k; y_i^k is the expression value of gene k in sample i; β_{ij} is the beta value of CpG j in sample i; and ϵ_i is a variable with normal distribution $\mathcal{N}(0, \sigma^2)$.

For each gene k we conduct a hypothesis test under the null hypothesis that $\gamma_j = 0$ for $j = 0, 1, 2, \ldots, m$. This is equivalent to stating that the beta values of the CpGs in a gene (independent variables) do not explain the expression level of the gene (dependent variable). On the other hand, the alternative hypothesis states that some $\gamma_j \neq 0$ and this implies that the beta values of some CpGs can explain the expression values of the gene in a linear fashion. P-values are obtained for each coefficient γ_j of gene k and for the gene itself. The fitness of the linear regression model for the gene is measured with an F-test. Genes with an F-test p-value < 0.05 are reported.

4.4.7 LASSO regression

Similarly to the linear regression model mentioned above, the LASSO method (least absolute shrinkage and selection operator) (Tibshirani, 2011) is used to conduct a regression analysis on beta values with respect to gene expression values. As shown in the next equation, if for gene k we have m differentially methylated sites, then the LASSO penalized regression can be formulated in Lagrangian form (Hastie et al., 2009) as in Equation 4.4:

$$\hat{\boldsymbol{\alpha}}^k = \underset{\boldsymbol{\alpha}^k}{\operatorname{argmin}}\{\frac{1}{2}\sum_{i=1}^{N}(y_i - \alpha_0 - \sum_{j=1}^{m}\beta_{ij}\alpha_j)^2 + \lambda\sum_{j=1}^{m}|\alpha_j|\}, \tag{4.4}$$

where $\hat{\boldsymbol{\alpha}}^k$ is the set of parameters we want to determine for gene k; $\lambda \geq 0$ is the penalty parameter to be determined based on cross-validation; and the rest of the parameters are the same as in multiple linear regression.

The intuition behind Equation 4.4 comes from understanding the two forces at play in the equation. On the one hand, the term $\frac{1}{2}\sum_{i=1}^{N}(y_i - \alpha_0 - \sum_{j=1}^{m}\beta_{ij}\alpha_j)^2$ wants to make use of all the coefficients $\boldsymbol{\alpha}^k = (\alpha_0, \alpha_1 ..., \alpha_m)$ in order to find a fit that minimizes the distance between the true expression value y_i and the predicted value $\alpha_0 + \sum_{j=1}^{m}\beta_{ij}\alpha_j$. On the other hand, the penalty term $\lambda\sum_{j=1}^{m}|\alpha_j|$ will attempt to truncate to zero as many αs as possible in order to minimize the entire equation. If the optimal solution to Equation 4.4 contains a coefficient $\alpha_j = 0$, it implies that the jth CpG and its beta values (β_{ij} with $i = 1, 2, \ldots, N$) are not strong predictors of the expression of the gene. Thus, the LASSO regression reduces the number of CpGs that is needed to consider for each gene.

4.4.8 Packages used

All the analyses and statistical tests implemented in `me-mRNA-pipe` were developed in R. The `limma` package (Smyth, 2004) is used for the differential expression/methylation analysis. The ANOVA analysis, Fisher's exact test and the multiple linear regression (linear fitting models) are based on functions in the `stats` package. The LASSO implementation is based on the `glmnet` package (Friedman et al., 2010).

4.5 Results for the Illumina 450K platform: a case study

We obtained a publicly available dataset from a study of how DNA methylation regulates lineage-specifying genes in the human vascular system (Brönneke et al., 2012). The study, which we will refer in this document as Brönneke's study, analyzed two different types of dermal cells:

- LEC: lymphatic endothelial cells (10 samples)
- BEC: blood endothelial cells (6 samples)

These types of cell were considered as the different phenotypes when running `me-mRNA-pipe`. The results presented in this section constitute an independent re-analysis of the data and are intended to show the output obtained from the pipeline when invoked on a dataset downloaded from GEO.

Brönneke's study (GEO Series GSE34487) jointly analyzed mRNA expression and methylation in LEC and BEC. For the analysis of mRNA expression profiles the Whole Human Genome Microarray 4×44K chip (from Agilent) was used. The analysis of the DNA methylation data was performed using the Illumina 450K array described in section 4.3.2.

During the execution of `me-mRNA-pipe`, a pre-processing step analyzes the methylation and mRNA data separately. This pre-processing step creates plots that are useful at assessing the quality of the data. The pre-processing is followed by an integrative analysis of the methylation and mRNA profiles. The genomic locations of the probes in each array –as defined by the manufacturer– are used to determine the association between mRNA transcripts and CpG dinucleotides. It is important to note that `me-mRNA-pipe` does not require the mRNA and methylation platforms to be from the same manufacturer and the analysis of Brönneke's data illustrates this principle.

The remainder of this section shows the results obtained by running `me-mRNA-pipe` on the datasets from Brönneke's study downloaded from GEO as Series Matrix files.

4.5.1 Data pre-processing

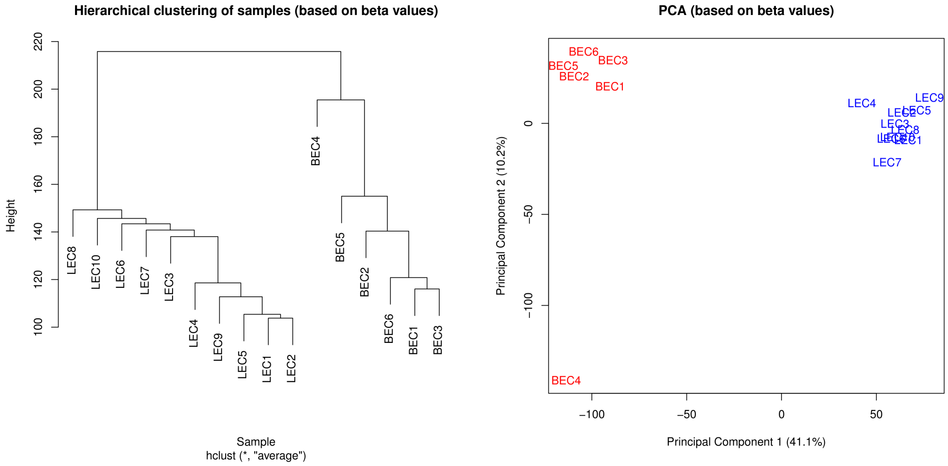
The pipeline runs a separate pre-processing on the mRNA and methylation samples as a quality control step. Firstly, a hierarchical clustering is performed on the samples followed by a principal component analysis (PCA). As annotated by Brönneke in GEO, the phenotypes of each sample are shown in Table XV. The column sample Id was manually added to link the mRNA and methylation samples.

The hierarchical clustering shows how closely the samples of each phenotype cluster together. The first two principal components of the PCA are used to plot the samples in two-dimensional space. Figure 25 illustrates this for the methylation data.

TABLE XV

PHENOTYPE INFORMATION FOR METHYLATION AND MRNA SAMPLES IN BRÖNNEKE *ET AL.* (GEO SERIES GSE34487).

Methylation data			**mRNA data**		
Illumina 450K array			Agilent Human Genome 4×44K array		
Sample Id	GEO Id	Phenotype	Sample Id	GEO Id	Phenotype
BEC1	GSM849975	BEC	BEC1	GSM812764	BEC
BEC2	GSM849977	BEC	BEC2	GSM812765	BEC
BEC3	GSM849988	BEC	BEC3	GSM812766	BEC
BEC4	GSM849989	BEC	BEC4	GSM812767	BEC
BEC5	GSM849979	BEC	BEC5	GSM812768	BEC
BEC6	GSM849984	BEC	BEC6	GSM812769	BEC
LEC1	GSM849974	LEC	LEC1	GSM812754	LEC
LEC2	GSM849976	LEC	LEC2	GSM812755	LEC
LEC3	GSM849980	LEC	LEC3	GSM812756	LEC
LEC4	GSM849978	LEC	LEC4	GSM812757	LEC
LEC5	GSM849983	LEC	LEC5	GSM812758	LEC
LEC6	GSM849981	LEC	LEC6	GSM812759	LEC
LEC7	GSM849982	LEC	LEC7	GSM812760	LEC
LEC8	GSM849986	LEC	LEC8	GSM812761	LEC
LEC9	GSM849985	LEC	LEC9	GSM812762	LEC
LEC10	GSM849987	LEC	LEC10	GSM812763	LEC



(a) Hierarchical clustering of samples

(b) PCA of samples

Figure 25. Pre-processing of methylation data

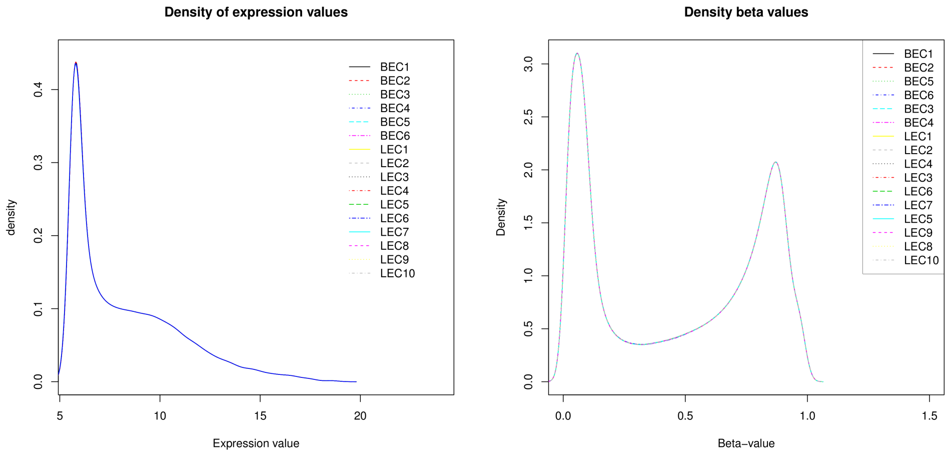
As it can be seen in Figure 25(a), the hierarchical clustering and PCA of the beta values show the samples grouped by phenotype as expected. These plots are useful in identifying a sample outlier. If, for example, a sample of phenotype 1 clusters closer to samples of phenotype 2, or if it is in a cluster of its own, this is normally a sign of experimental errors and that it may be necessary to remove the sample from further analysis.

The first principal component of the PCA in Figure 25(b) explains 41% of the variation in the data and is enough to discriminate the BEC (red) from LEC (blue) samples. The second principal component explains only 10% of the variation and it does not seem to differentiate samples in LEC. On the other hand, the second principal component shows one sample of BEC as a potential outlier (GEO id = GSM849989). Because we are running this analysis for illustration purposes in this reanalysis all samples were included.

Another important quality assessment that needs to be performed on the data is to guarantee the samples are normalized. Generally, datasets of studies downloaded from GEO were normalized by their authors. The pipeline generates density plots of both mRNA expression values and methylation beta values. Figure 26 shows the density of mRNA expression values (a) in mRNA samples and of CpG values (b) in methylation samples.

From Figure 26 we can infer that:

- all samples are normalized since their curves have almost identical overlap
- in (b), the bimodal nature of the curves shows that there are more CpGs with beta values close to zero than there are close to one.



(a) mRNA expression

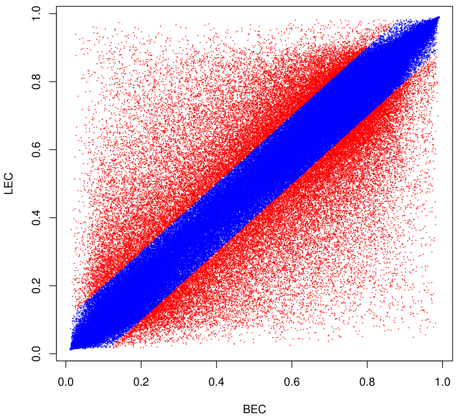
(b) Methylation, beta values

Figure 26. Density of mRNA and methylation probes in Brönneke's dataset.

In our reanalysis, mRNA probes were considered to be differentially expressed if the log2-fold change between phenotypes was greater than 0.6. This is equivalent to more than a 1.5-fold difference in mRNA levels between phenotypes. Similarly, DNA methylation probes were considered to be differentially methylated if the difference in beta values (Δbeta) was greater than 0.1. The pipeline provides other methods to compute differential expression/methylation but the best results will be obtained when the user obtains the original lists of differentially methylated CpGs and differentially expressed genes. Figure 27(a) shows a summary of the number of differentially expressed mRNAs, the number of differentially methylated CpGs and their overlap computed by `me-mRNA-pipe`. A scatter plot of all the beta values in both phenotypes is shown in Figure 27(b).

	Count
Differentially Methylated CpGs (DMCpG)	50,437
Differentially Expressed RefSeqIds (DER)	1,574
▷ DMCpGs that overlap with at least one DER	5,285
▷ DERs that overlap with at least one DMCpG	1,173

(a) Summary of differentially expressed mRNAs and CpGs



(b) Beta values of CpGs in both phenotypes. Differentially methylated CpGs are colored in red.

Figure 27. Differential expression/methylation.

4.5.2 ANOVA analysis

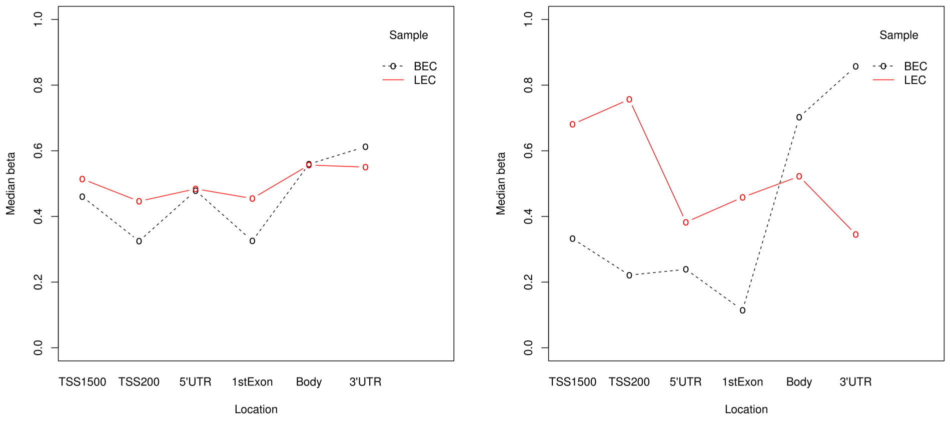
The integrative portion of the analysis is based on differentially expressed genes that are associated, i.e., overlap with differentially methylated CpGs. In particular, the ANOVA analysis as described in Equation 4.2 attempts to determine if there is a statistically significant interaction between the location of a CpG and the phenotype.

When applying Equation 4.2 to Brönneke's study we have β^k_{pq} as the beta value of a CpG associated to gene k; ϕ_p is the phenotype effect for phenotype $p = \{\text{BEC, LEC}\}$, τ_q is the location effect for location $q = \{\text{TSS1500, TSS200, 5'UTR, 1stExon, Body, 3'UTR}\}$ when considering the possible values of location by gene as defined in the Illumina 450K array, and

$(\phi\tau)_{pq}$ is the interaction effect between phenotype and location for every possible pair $pq = \{$(TSS1500, BEC), (TSS1500, LEC), (TSS200, BEC), …, (3'UTR, LEC)$\}$.

The ANOVA tests are conducted on a per-gene basis and they rank all genes based on the p-values obtained from the interaction test $(\phi\tau)_{pq}$. The top K genes and bottom K genes are reported to the user. Because we only analyze genes that are differentially expressed, both the top and bottom sets are significant and deserve attention. On the one hand, the set of top K genes is comprised by genes whose overlapping CpGs have a statistically significant interaction between the phenotype effect and location effect. On the other hand, the bottom K are genes where this interaction is very unlikely to be present (p-value $\simeq 1.0$) and methylation patterns are consistent across different locations in the genes. Brönneke's dataset was analyzed with $K = 20$.

Figure 28 shows the contrast in median beta values of CpGs that overlap with all differentially expressed genes and with the set of top 20 genes with smallest p-values reported by our test (9.2e-155 $<$ p-value $<$ 2.1e-18). In Figure 28(a) we see that beta values in LEC are equal or slightly larger than the beta values in BEC, for all locations except for the 3'UTR region. When we focus on the top 20 genes and their CpGs, Figure 28(b), we can see that the methylation levels of LEC are markedly higher in the promoter region and up to the first exon. In particular, methylation levels in BEC seem to decrease in the proximity of the TSS whereas in LEC we see an increase from the distal promoter (TSS1500) to the vicinity of the TSS (TSS200).



(a) All differentially expressed genes (b) Top 20 genes ranked by ANOVA

Figure 28. Median CpG beta value per phenotype at different locations

In search for a biological interpretation of these interactions, we conducted GO term and pathway enrichment analysis using an external tool. The list of top 20 genes output by the pipeline was submitted to DAVID (Database for Annotation, Visualization and Integrated Discovery) (Huang et al., 2008). We identified that the proteins encoded by these genes have a statistically significant enrichment of Pleckstrin homology (PH) domains in two different databases: InterPro (Hunter et al., 2012) and SMART (Schultz et al., 1998). PH domains are known to recruit proteins to different membranes, thus mediating protein-phospholipid interactions and interacting with signal transduction pathways. The results of the enrichment are shown in Table XVI.

TABLE XVI

FUNCTIONAL ANNOTATION OF TOP 20 GENES FROM THE ANOVA TEST.

Id	**Term**	**Category**	***p*-value**
IPR011993	Pleckstrin homology-like domain	InterPro	5.0e-3
SM00233	PH, Pleckstrin homology domain	SMART	6.9e-3

The table shows terms with a false discovery rate (FDR) of less than 20%. There are multiple other GO terms with an enrichment *p*-value < 0.05 that did not meet our FDR criterion but which are meaningful and may accentuate the importance of this subset of genes. Some of these terms are: GO:0040008-regulation of growth ($p = 1.1$e-2); GO:0042981-regulation of apoptosis ($p = 2.1$e-2) and GO:0048538-thymus development ($p = 3.0$e-2).

Finally, to stress the importance of the ANOVA analysis, we want to relate our results to the ones obtained by Brönneke *et al.* In their work, the authors used Ingenuity Pathway Analysis to build a top-score network that contained 24 genes. Of these 24 genes, 5 of them were identified by our ANOVA analysis as having interactions between phenotypes and CpG locations (*p*-value < 0.05), and 2 of them are in our list of top 20 genes. The genes are: AEBP1, ELK3, IL7, PROX1 and TBX1 (with two statiscally significant transcripts NM_080646 and NM_080647). It is important to note that these 5 genes were selected solely because of interactions present in the CpGs associated to them. So far we have not used gene expression data in any meaningful way. The following sections address this issue.

4.5.3 Correlation analysis

This analysis is the first step at integrating gene expression and methylation data. Here we consider the correlation of beta values with respect to gene expression. As described in section 4.4.5, we have N expression values $[y_1, y_2, \ldots, y_N]$ for gene k. Similarly, for the jth CpG that overlaps with gene k we will have N beta values $[\beta_{1,j}, \beta_{2,j}, \ldots, \beta_{N,j}]$ corresponding to matching methylation samples. In Brönneke's dataset N=16 and for each gene k and CpG$_j$ we compute a correlation coefficient ρ_j^k.

When we group these correlation coefficients by the location of the CpG we can determine if there is a statistically significant difference between coefficients in one location versus the other. We can phrase the previous statement as a question: "Are there more positively/negatively correlated CpGs at location *loc* than in the rest of the locations?". In order to answer this, the pipeline creates contingency matrices like the one shown in Figure 24 and then runs a Fisher's exact test for each location.

To add one more level of complexity we have three types of locations, i.e.: location by gene, location by CpG island and custom location, each of them with different location values (see sections 4.3.2 , 4.3.3 and 4.3.4). By repeating Fisher's exact test on these locations, we can determine which type of location provides a richer view of how methylation correlates with gene expression. Figure 29 shows the results of these tests in Brönneke's dataset.

It is clear that the location by CpG island in Figure 29(b) and its more compact version, the custom location, in Figure 29(c), have few locations where there are statistically significant differences in correlation coefficients. On the other hand, when looking at the location by

Location	***p*-value**
TSS1500	0.001034
TSS200	0.000011
5'UTR	0.000036
1stExon	~ 0.0
Body	~ 0.0
3'UTR	0.000009

(a) Location by gene

Location	***p*-value**
N_Shelf	0.731691
N_Shore	0.912680
Island	0.760824
S_Shore	0.001777
S_Shelf	0.259031
Open_sea	0.019594

(b) Location by CpG island

Location	***p*-value**
Island	0.760824
Shore	0.008399
Open_sea	0.019594

(c) Custom location

Figure 29. Results of Fisher's exact test for different types of locations. Significance level $\alpha = 0.01$

gene, Figure 29(a), we see that all locations are significant. The importance of this finding cannot be overstated. One of our goals in developing `me-mRNA-pipe` is to help shift the focus of methylation analysis from a CpG island-centered context to a gene-oriented perspective. In this scenario, individually methylated CpGs –distant from an island– can still have an effect on gene expression if they are located in the *right* regions of a gene.

Knowing of the importance of the location by gene, we can create density plots of correlation coefficients at each of its locations (Figure 30).

From the figure we observe that all locations have a larger number of negatively correlated CpGs, except for the 3'UTR region. The curve for 3'UTR (magenta-pink) shows a higher density of positively correlated CpGs. It is important to note that our results about the positive correlation of CpGs in the 3'UTR region coincide with findings reported by Brönneke *et al.* The authors mention in their paper: *"...3'UTR appeared to be hypermethylated in upregulated genes and hypomethylated in downregulated genes."* They continue to explain why this finding in the

3'UTR region is important: *"...It has been proposed that methylation in the 3'UTR may play a role in suppression of antisense transcripts, regulation of polyadenylation, and termination of transcription, respectively..."*

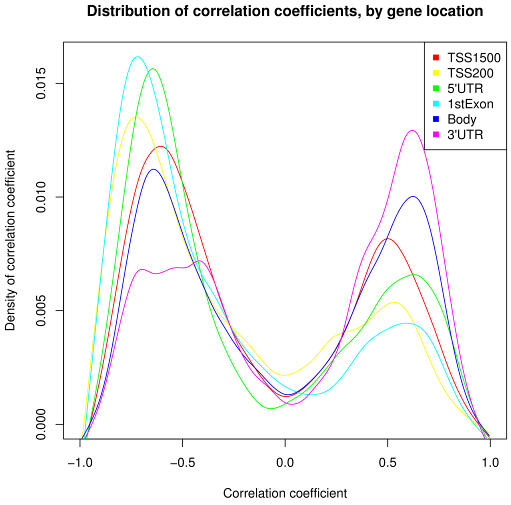


Figure 30. Density of correlation coefficients (location by gene).

This is an important validation of the results obtained by `me-mRNA-pipe`, especially since we conducted an independent reanalysis of Brönneke's data. The next two sections provide details of the regression analysis implemented in the pipeline, which we believe are the most novel and salient features offered by `me-mRNA-pipe`.

4.5.4 Multiple linear regression

With the ANOVA analysis we identified important genes simply by analyzing the variation of the beta values of overlapping CpGs. Important as this is, the analysis is limited to genes that have overlapping CpGs at different locations. For example, one of the differentially expressed genes in our analysis is LAMA5 (laminin, alpha 5). Laminins are a family of extracellular glycoproteins that are hypothesized to ...“*mediate the attachment, migration, and organization of cells into tissues during embryonic development by interacting with other extracellular matrix components...*” (Maglott et al., 2005). In the context of our analysis, LAMA5 is an important gene which, in addition to being differentially expressed, has four differentially methylated CpGs that overlap with its coding region. Due to the fact that these four CpGs are located in the same location (location by gene = Body), the ANOVA analysis is unable to identify any interaction between location and phenotype. To address this issue, we decided to explore the relationship between the expression values of genes and the beta values of the CpGs that overlap with them using regression analysis as described in section 4.4.6. Here we are interested in modeling the expression of genes with a linear combination of the beta values of their overlapping CpGs. The model was described in Equation 4.3.

Figure 31 shows the first 35Kb of LAMA5 (RefSeq Id: NM_005560) obtained from the UCSC Genome Browser (Kent et al., 2002; Fujita et al., 2010). The figure also shows the location of its four overlapping CpGs. LAMA5 is 58,248 bp long, located in chr20 complement(60,884,121-60,942,368) and therefore transcribed in the 3’-5’ direction.

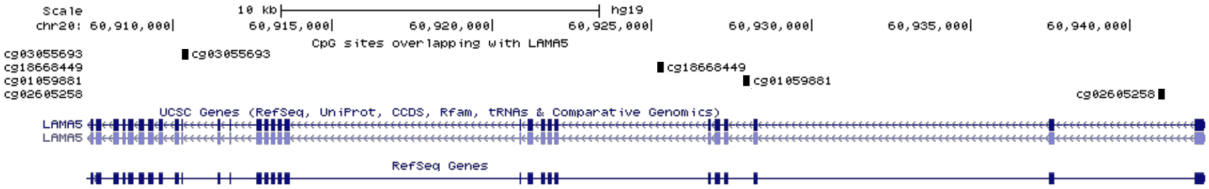


Figure 31. Genomic location of LAMA5 (NM_005560) and its overlapping CpGs.

As an example, we will illustrate the steps `me-mRNA-pipe` takes to determine if Equation 4.3 is a reliable model for the gene and its CpGs. Table XVII contains the expression values of the gene and the beta values of the CpGs across all samples (the CpGs are listed, from left to right, in order of their proximity to the TSS).

We can infer from Table XVII that LAMA5 is over-expressed in BEC (the average expression value in BEC samples is greater than the average expression value in LEC, $10.37 > 8.99$ respectively). When considering the average beta values per phenotype, we have cg01059881 and cg03055693 hyper-methylated in BEC whereas cg02605258 and cg18668449 are hypo-methylated in BEC. The question we want to address is: Do these CpGs contribute statistically to the observed difference in expression levels of LAMA5 between BEC and LEC? And if they do, *how* can we measure their contribution?

`me-mRNA-pipe` conducts a multiple regression analysis on all differentially expressed genes and the differentially expressed CpGs with which they overlap. For each gene, the input to the

TABLE XVII

EXPRESSION VALUES OF LAMA5 (NM_005560) AND BETA VALUES OF ITS ASSOCIATED CPGS IN ALL SAMPLES.

	y_i^{LAMA5}	β_{ij}			
Sample Id	**Gene expression**	**cg02605258**	**cg01059881**	**cg18668449**	**cg03055693**
BEC1	10.247423	0.674306	0.574543	0.553044	0.778331
BEC2	10.493272	0.734266	0.726371	0.590922	0.694452
BEC3	10.238017	0.492988	0.763868	0.527128	0.846899
BEC4	10.601016	0.616107	0.642293	0.618739	0.798540
BEC5	10.322472	0.705888	0.584753	0.643651	0.804295
BEC6	10.297545	0.769251	0.585207	0.727116	0.866327
LEC1	9.039718	0.892263	0.501706	0.766910	0.667069
LEC2	9.001532	0.861426	0.466992	0.755265	0.664860
LEC3	8.795813	0.797434	0.581737	0.726024	0.648924
LEC4	9.912273	0.843690	0.412705	0.787442	0.771526
LEC5	9.720359	0.847242	0.361803	0.757306	0.758610
LEC6	8.972494	0.849028	0.519599	0.756435	0.743085
LEC7	8.628658	0.847183	0.534746	0.737375	0.529687
LEC8	8.753028	0.834574	0.533669	0.766279	0.631078
LEC9	8.368857	0.838693	0.398298	0.762841	0.679842
LEC10	8.671963	0.849822	0.462896	0.730885	0.714444

analysis is a table similar to Table XVII. In the case of LAMA5, the p-value obtained from the multiple regression linear analysis was statistically significant (F-statistic = 8.173 and p = 2.6e-3). The intercept and coefficients of the CpGs are shown in Equation 4.5 and Table XVIII.

$$y_i^{\text{LAMA5}} = 5.71 + 2.70\ \text{cg02605258}_i + 1.66\ \text{cg01059881}_i - 5.05\ \text{cg18668449}_i + 6.00\ \text{cg03055693}_i \tag{4.5}$$

TABLE XVIII

COEFFICIENTS AND P-VALUES OF MULTIPLE LINEAR REGRESSION FOR LAMA5.

	Coefficients		***t*-statistic**	
	Name	**Value**	**Value *t***	***p*-value**
γ_0	intercept	5.7050	1.490	0.1644
γ_1	cg02605258	2.6998	0.816	0.4318
γ_2	cg01059881	1.6550	0.823	0.4280
γ_3	cg18668449	-5.0545	-1.364	0.1999
γ_4	cg03055693	5.9915	3.169	0.0089

where y_i^{LAMA5} is the expression value of LAMA5 in sample i and the cg_id_i represents the beta value of the CpG in sample i. It is important to note that the final goal of this analysis is not to obtain an equation like Equation 4.5. We do not want to predict the expression values of LAMA5. Nonetheless, we want to know if such an equation is valid and, through the coefficients in the equation, we can infer interesting properties of the CpGs that overlap with the gene. For example:

- From Table XVII, we inferred that LAMA5 is over-expressed in BEC
- CpGs cg01059881 and cg03055693 are hyper-methylated in BEC. Their coefficients in Equation 4.5 are positive, therefore the higher their methylation in BEC, the higher the expression of LAMA5.
- CpGs cg02605258 and cg18668449 are hypo-methylated in BEC. cg02605258 has a positive coefficient whereas cg18668449 has a negative coefficient. This seems conflicting at first

and we cannot arrive at any conclusion. In the following section –using LASSO regression– this apparent contradiction is resolved.

The multiple linear regression analysis is conducted on a gene-by-gene basis and `me-mRNA-pipe` creates output results in the form of plots and text files for each gene. The coefficients and p-values in Table XVIII were obtained from one of these text files. Among other results, there is a group of regression diagnostic plots that are worth mentioning. These regression plots inform the user about potential violations of the model assumptions, for example: when errors have unequal variance, and/or are non-normally distributed. They can also provide information about possible outliers.

To conclude this analysis, we would like to again compare our results to the ones reported by Brönneke *et al.* Of the 24 genes the authors reported in their top-score network, 16 of them have regression p-values (from F-test) less than 0.05. This is important as it stresses the point that for the majority of the relevant genes reported by Brönneke *et al.* the linear regression between beta values and gene expression values is statistically significant. Table XIX lists the 16 genes with their respective p-values.

4.5.5 LASSO regression

Through multiple linear regression we were able to identify genes with overlapping CpGs whose beta values are good descriptors of mRNA expression values. Yet, as it was shown for LAMA5, sometimes we can find CpGs whose regression coefficients convey conflicting information. Therefore, we want an answer to the following question: “If the differentially expressed

TABLE XIX

GENES FROM TOP-SCORE NETWORK IN BRÖNNEKE *ET AL.* WHOSE REGRESSION IS STATISTICALLY SIGNIFICANT.

Gene symbol	**RefSeq Id**	**Number of associated CpGs**	**Regression F-statistic p-value**
AEBP1	NM_001129	3	2.2e-4
BATF	NM_006399	2	8.6e-4
CD36	NM_001001547	4	2.9e-4
ELK3	NM_005230	4	2.9e-2
FABP5	NM_001444	2	6.6e-4
IL7	NM_000880	7	4.4e-3
ITGA10	NM_003637	1	3.9e-4
MAF	NM_001031804	2	~ 0.0
MRC2	NM_006039	3	2.0e-2
NID1	NM_002508	4	5.2e-3
PROX1	NM_002763	11	8.3e-3
RELN	NM_005045	6	5.3e-3
RTKN	NM_033046	5	8.0e-3
SLC2A12	NM_145176	3	6.5e-4
TFPI	NM_001032281	4	1.6e-4
TFPI	NM_006287	4	2.6e-3
VAV3	NM_006113	5	2.5e-3

gene k overlaps with m differentially methylated CpGs, which group of $s < m$ CpGs can be prioritized and still be good predictors of the expression of gene k?"

In a generic way, Figure 32 illustrates this problem. We will have n differentially expressed genes, from $gene_1$ through $gene_n$. These genes will be over- or under-expressed in one phenotype (BEC in our case). For gene k, some of the CpGs will be hypo-methylated and others will be hyper-methylated. By prioritizing these CpGs using LASSO regression, i.e., keeping the CpGs that best predict the expression of the gene, we not only reduce the complexity of the problem, but also obtain a better picture of the regions in the gene where methylation has a stronger effect in differential expression.

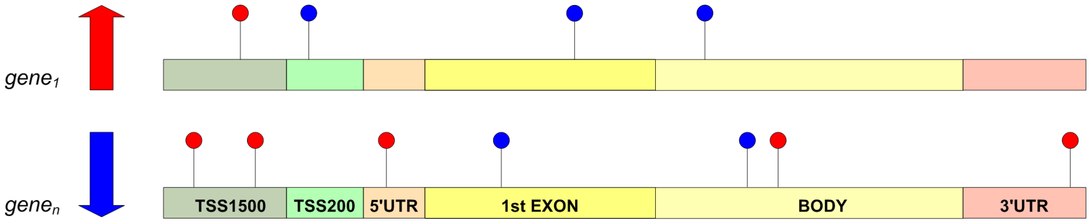


Figure 32. Differentially methylated sites associated to differentially expressed genes. Blue/red arrow: down/up-regulated gene in phenotype 1 compared to phenotype 2; blue/red methylation site: hypo/hyper-methylated.

For Brönneke's data, and as it was the case for multiple linear regression, the y_i and β_{ij} values are like the ones described in Table XVII. In fact, when we apply LASSO regression on the gene LAMA5 we obtain the solution listed in Table XX. In order to obtain the values in Table XX, we have to solve Equation 4.4 in iterative steps:

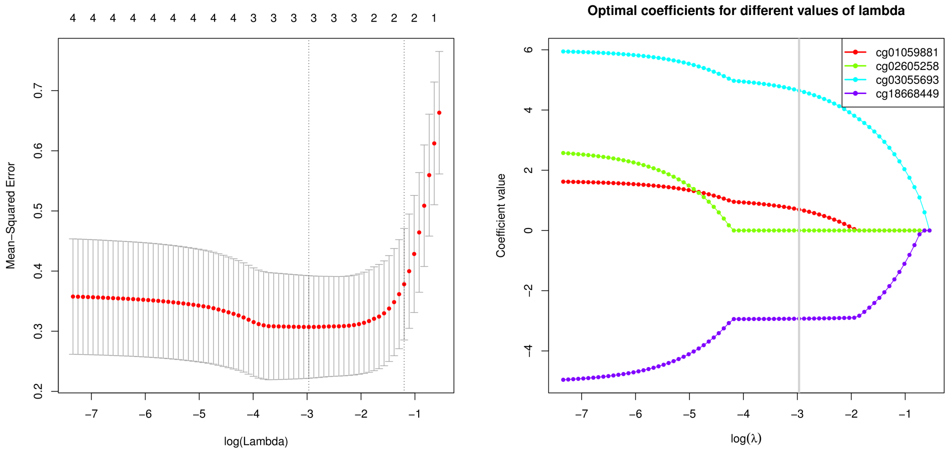
- Estimate a value for λ
- Minimize the equation using the estimated λ

TABLE XX

SOLUTION TO LASSO REGRESSION FOR LAMA5.

	Parameters		**Value**
λ	Obtained from cross-validation		0.0514
$\hat{\boldsymbol{\alpha}}^k$	α_0	intercept	7.8129
	α_1	cg02605258	0.0000
	α_2	cg01059881	0.6961
	α_3	cg18668449	-2.9290
	α_4	cg03055693	4.6441

This iteration is performed by doing leave-one-out cross-validation until a set of values $\{\lambda,\ \hat{\boldsymbol{\alpha}}^k\}$ that minimizes the equation is found. Figure 33(a), with natural log values of λ in the X-axis, shows the iterative process and the number of coefficients (top of the figure) that are selected by LASSO at each λ. The λ that minimizes the mean cross-validated error is 0.0514 and is marked by a vertical line at $ln(0.0514) = -2.9681$. The second vertical line in Figure 33(a) is at $ln(\lambda) = ln(0.3010) = -1.2006$ and indicates the largest value of λ that yields a mean cross-validated error less than one standard deviation from the minimum. On the other hand, Figure 33(b) marks the values the coefficients take for the λ that minimizes the mean cross-validated error with a vertical line at $\lambda = 0.0514$. It is clear from this figure that as $\lambda \to 0$ the penalty term in Equation 4.4 becomes zero and we obtain a standard solution to multiple linear regression (see the values the coefficients take in the Y-axis, in the left part of the Figure 33(b), and compare them to Equation 4.5). Conversely, as $\lambda \to 1$ the penalty term puts a higher weight on the equation and all coefficients are set to zero. The key to this method is to strike a balance and to find a subset of the CpGs, not too large and not too small, that best explains the expression levels of the gene.



(a) Number of coefficients chosen at each λ value (leave-one-out cross-validation)

(b) Values of coefficients –not including intercept– after minimization at each λ

Figure 33. Iterative minimization of LASSO equation for LAMA5.

When multiple linear regression was applied on LAMA5, we found that cg02605258 and cg18668449 have regression coefficients with opposite signs despite both of them being hypomethylated in BEC. Nevertheless, in Table XX we can see that LASSO regression assigned a zero coefficient to cg02605258. The final model "dropped" cg02605258 and kept the three other CpGs. The coefficients for the remaining CpGs have the same sign as in multiple linear regression and they provide the following interpretation of the results (extended from multiple linear regression):

- We know LAMA5 is over-expressed in BEC
- We also know that cg01059881 and cg03055693 are hyper-methylated in BEC. Their positive coefficients in Table XX imply that higher levels of methylation in BEC produce higher expression of LAMA5.
- From our LASSO results, we exclude cg02605258 from subsequent analysis.
- Finally, we know cg18668449 is hypo-methylated in BEC. The negative coefficient in cg18668449 means that a lower level of methylation in that position begets a higher expression of LAMA5.

Figure 34 shows the strong correlation (Pearson's correlation coefficient $\rho = 0.856$) between the known expression values of LAMA5 (column y_i^{LAMA5} in Table XVII) and the predicted values using the coefficients obtained from LASSO in Table XX.

To complement the analysis of Figure 34 and for illustration purposes, we decided to re-run multiple linear regression of LAMA5 but only on the coefficients selected by the LASSO method (this filtered re-execution is not currently supported by the pipeline). The results shown in Table XXI can now be compared to the original results in Table XVIII. Without cg02605258, the linear model for LAMA5 has a much better fit with a more significant p-value (F-statistic $= 10.98$ and $p = 9.3$e-4).

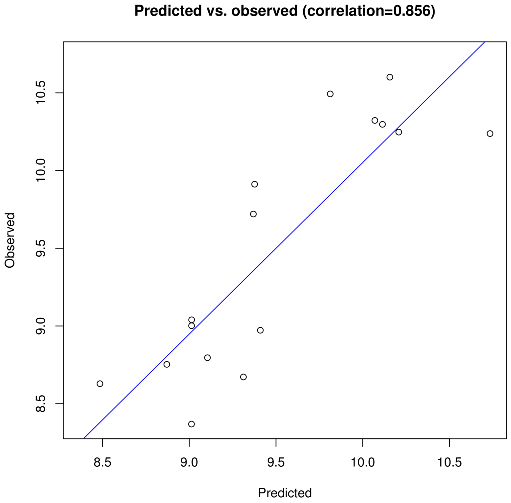


Figure 34. Predicted vs. observed values for LAMA5. Prediction based on the three coefficients obtained from LASSO regression.

TABLE XXI

COEFFICIENTS AND P-VALUES OF MULTIPLE LINEAR REGRESSION FOR LAMA5 AFTER FILTERING CPG WITH LASSO.

	Coefficients		***t*-statistic**	
	Name	**Value**	**Value *t***	***p*-value**
γ_0	intercept	7.2853	2.236	0.0451
γ_1	cg01059881	1.0631	0.575	0.5760
γ_2	cg18668449	-2.9439	-1.125	0.2827
γ_3	cg03055693	5.1127	3.336	0.0059

In conclusion, the LASSO penalized regression method is useful at identifying CpGs that contribute poorly to the differences in mRNA expression levels. After excluding these CpGs from further analysis we have shown that the predictability of the linear model increases significantly. This allows the user to focus on a smaller set of CpGs, which with additional experimentation, can be tested for their role in regulation of gene expression.

4.6 Conclusion

We have developed `me-mRNA-pipe`, a software tool that integrates the analysis of mRNA and methylation microarray data. The execution of the pipeline is fully customizable through configuration files and creates graphic plots, summary statistics and tab-separated text reports that can be easily imported into spreadsheets.

We expect that the output generated by `me-mRNA-pipe` will provide the user with an encompassing view of how methylation affects gene expression. The pipeline uncovers statistically significant interactions of beta values between phenotypes and CpG locations, in addition to exploring the correlations between CpGs at specific locations and gene expression. The regression analysis assumes a linear model of beta values with gene expression and therefore reports those genes where this relationship holds with statistical significance. Finally, the LASSO penalized regression allows the user to focus on a smaller set of CpGs per gene –and their locations with respect to the gene– to further analyze the potentially regulatory role of these CpGs. [1]

[1] Refer to the Appendix to see the Results for the Illumina 27K platform.

CHAPTER 5

A NORMALIZATION PROCEDURE FOR THE ANALYSIS OF CHROMOSOME CONFORMATION DATA.

5.1 Introduction

As it was shown in previous chapters, our goal has been to develop methods that will help us gain a better understanding of how genes are regulated. Therefore, it seems of utmost importance to complete this journey by delving into the complex interplay between chromatin conformation and gene regulation. Chromosomes show a non-random spatial organization in the nucleus. These conformations act as scaffolds that regulate genome functions and epigenetic inheritance in different cell states. There are well documented cases of enhancers that affect the expression of distantly located genes, sometimes in different chromosomes. This is achieved through direct interaction of the enhancer and the promoter of the target gene (West and Fraser, 2005). With this in mind, we can view the genome as a three-dimensional entity where chromosomes occupy certain "territories" and gene regulation is affected by physical interactions between genes and regulatory elements located in the same territory.

This chapter addresses all the technical hurdles we faced and the methods we developed in order to obtain a reliable picture of how chromosomes fold under specific conditions.

5.2 Preliminaries

From a historical perspective, the first high-throughput technology used for the analysis of chromosome conformation was “Chromosome Conformation Capture”, known as 3C (Dekker et al., 2002). Prior to the advent of 3C, microscopy-based technologies such as the fluorescence in-situ hybridization method (FISH) have been used to identify interactions between different genomic loci. 3C has the advantage that it can reveal these genomic interactions at a very high resolution without the adverse effects that FISH treatment could have on chromosome structure (Dekker et al., 2002). The goal of 3C is to find segments of DNA that are physically in contact with each other through DNA-bound proteins. This is achieved by subjecting nuclei to a crosslinking reagent, such as formaldehyde, which fixes proteins interacting with other proteins and with DNA. These fixed interactions have to undergo four more chemical reactions in order to be quantified: i) they are first digested with a restriction enzyme, ii) followed by a process of ligation, iii) then the cross-linking is reversed and finally iv) quantification of ligated products is done through PCR.

Over the years, 3C was improved with other technologies known as 4C (Circular Chromosome Conformation Capture) (Zhao et al., 2006) and 5C (Chromosome Conformation Capture Carbon Copy) (Dostie et al., 2006). These techniques are very effective at identifying interactions between pre-determined genomic segments, where by pre-determined we mean that the set of loci to investigate must be chosen in advance. This limitation was overcome with the introduction of Hi-C (Lieberman-Aiden et al., 2009) which can identify chromatin interactions in a genome-wide scale. Even if Hi-C does not require an a priori specification of the genomic region of interest, a technique like 5C is extensively used when the goal is to investigate a specific genomic region at a very high resolution. Because the methods in the rest of this chapter focus on the 5C technology, its main steps are illustrated in Figure 35. Please refer to the Appendix, Figure 65 for details about Hi-C.

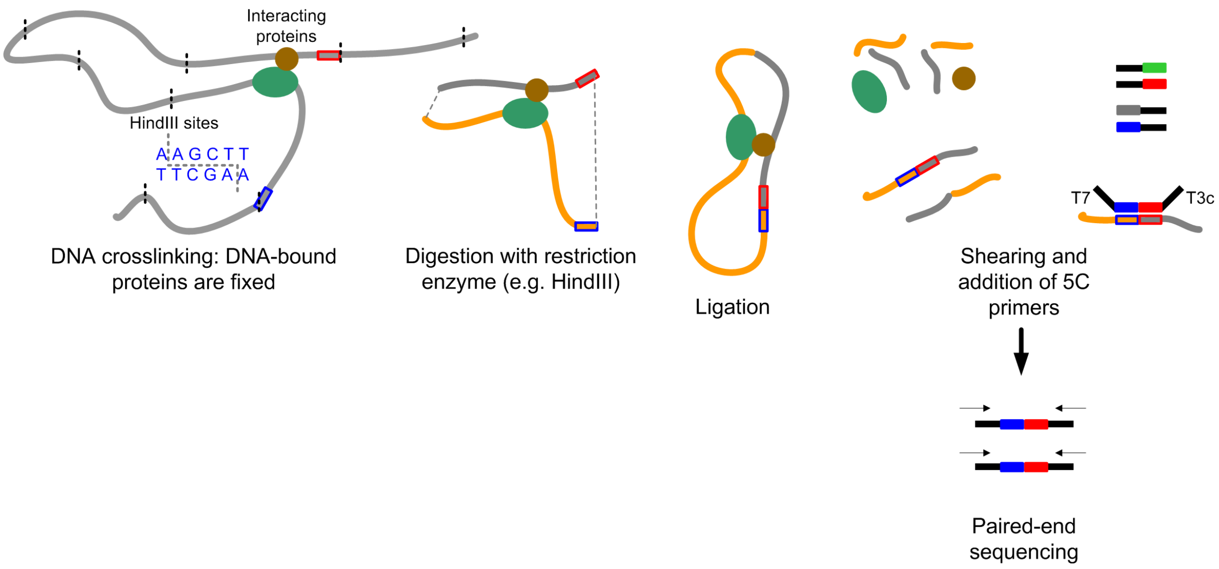

Interacting proteins
HindIII sites
AAGCTT
TTCGAA
DNA crosslinking: DNA-bound proteins are fixed
Digestion with restriction enzyme (e.g. HindIII)
Ligation
Shearing and addition of 5C primers
T7
T3c
Paired-end sequencing

Figure 35. An overview of the steps in 5C.

For this project, we will use the mouse as our model organism. We are particularly interested in finding the DNA interactions that occur within one of the three loci that are responsible for the production of immunoglobulins in B cells. Immunoglobulins (Ig) are highly specialized proteins that are used to recognize antigens from bacteria, viruses and other disease-causing organisms (Murphy, 2011). B cells generate immunoglobulins for a large range of antigen specificities and their production requires well-orchestrated DNA recombinations within very specific genomic regions. A group of helper proteins facilitates the recombination process by bringing sections of DNA close to each other and thus creating DNA interactions. Our goal is to detect these interactions and to describe what genomic locations are more likely to interact with each other. Of the three loci where immunoglobulins are produced –listed in Table XXII– we will focus on Igh between the coordinates listed in the last two columns of the table. Collectively, immunoglobulins, are also known as *antibodies* and in the rest of this chapter the terms will be used interchangeably.

TABLE XXII

IMMUNOGLOBULIN LOCI COORDINATES IN MOUSE.

			NCBI definition		Coordinates of our study	
Locus	Chr	Entrez ID	Start	End	Start	End
Igl (λ)	16	111519	19,026,858	19,260,844	–	–
Igk (κ)	6	111507	67,555,636	70,726,754	–	–
Igh (Heavy)	12	243469	113,258,768	116,009,954	114,353,686	117,349,200

5.3 Characteristics of the 5C data to analyze interactions in Igh locus

In addition to reliably quantifying the interactions that take place in the Igh locus, we wanted to develop a methodology that will allow us to compare interactions detected in different cell types. To that effect, we used experimental data obtained by Dr. Amy Kenter's Lab in the Department of Microbiology and Immunology at the University of Illinois-Chicago (publication is in preparation). The experiments focused on analyzing long range DNA interactions in two cell types: pro-B cells and mouse embryonic fibroblasts (MEF).

- pro-B cells: B-lymphocytes (B cells) produce immunoglobulins in response to an antigen. Pro-B cells, in turn, are B cells at the earliest stage in the cell's life cycle (Murphy, 2011). The reason why pro-B cells are interesting is because they undergo intense gene rearrangement necessary to produce immunoglobulins. And, needless to say, this rearrangement requires DNA interactions to take place at the specific sites where the rearrangement will occur.
- MEF cells: Fibroblasts are a type of cell that forms connecting tissue in animals. Fibroblasts are dispersed throughout the body and, in response to an injury, they proliferate and secrete a collagenous extracellular matrix rich in type I and II collagen to repair the wound (Alberts et al., 2002).

Because pro-B cells are involved in the production of immunoglobulins and MEF cells are not, by quantifying the interactions in the Igh locus for both cells we expect to determine what DNA interactions are pro-B specific and, thus, gain a better insight into the genomic regions within Igh that are involved in gene rearrangement.

The experimental data at our disposal consists of the following:

- pro-B: two biological replicates. For each replicate, an independent 5C experiment was conducted using a 5C primer design that probes interactions in two different locations:
  - Igh locus: our region of interest.
  - Gene desert in chromosome 5: used as an internal control.
- MEF: also two biological replicates. For each replicate, the primer design is the same as in pro-B, covering the same genomic regions:
  - Igh locus
  - Gene desert in chromosome 5, also used as internal control.

As it will become clear later in Section 5.7.2, the gene desert will be important in determining a scaling factor that will enable us to compare the different cell types. The following sections explain the type of data we had to handle and the technical steps we had to follow to be able to reliably quantify interactions.

5.4 Pre-processing of raw 5C data

The first step before quantifying interactions is the pre-processing of the raw sequencing data. If you recall in Figure 35, the output from 5C is a dataset with paired-end reads that need to be sequenced. After the data are sequenced, they constitute the input to our pipeline and, as with any ordinary dataset, it requires pre-processing. The pre-processing stage is dominated by the mapping of **paired-end reads** –obtained from the sequencer– to a reference genome.

Because all our experimental data were obtained from mouse cells, our reference genome is the mouse genome assembly mm9, July 2007, NCBI Build 37.

The *reads* are normally referred to as next-generation sequencing (NGS) data. In order to better understand the complexity and technicalities involved in processing reads, we will go step-by-step addressing the following questions:

- What are the characteristics of a raw read in an NGS file?
- How is the quality of a read measured? And how is this quality measure used?
- Are there any areas in the Igh locus where reads cannot be mapped reliably?
- If the answer to the previous question is yes, can we create a “virtual genome” to improve the chances of mapping all our reads?

5.4.1 Next-generation sequencing data

The sequencing that takes place is called paired-end sequencing and consists of sequencing each side of an interaction independently of the other (in the direction of the arrows shown in Figure 35). After they are sequenced, the two *reads*, one for each end, are stored in different files and the pairing is maintained by a unique identification the sequencer gives to them. Each of these files will have a format similar to the one described in Figure 36.

```
@HWI-EAS348:12:FC:1:1:2201:3496 1:N:0:
TAATACGACTCACTATAGCCATAAATTACTGAAGGGCTTAAG
+
GGGGGFHHHGHEHGBGGGGGH@HHGHH<GHBGG@GFEHHHHD

@HWI-EAS348:12:FC:1:1:2201:11118 1:N:0:
TATTAACCCTCACTAAAGGGAGCACAGCCTGATAAACACCTT
+
GGGGGGGGGGGGGGGGGGGGG8GG2-;/843;.>9EE#####
```

Figure 36. Two sample reads in a `FASTQ`.

The format of this file is known as `FASTQ` (FASTA sequence and its Quality scores). Normally, the information about a read is grouped in 4 consecutive lines, where each line represents:

- **Line 1**: Starts with the `'@'` character, followed by a unique identifier created by the sequencer. In the case of paired-end reads, the last number in the identifier will be different. For example, let's assume the file `p1.fastq` contains all the reads of the first pair in paired-end sequencing. File `p2.fastq` contains all the reads of the second pair. If the identifier of a read in `p1.fastq` is `@HWI-EAS348:12:FC:1:1:2201:3496 1:N:0:` as shown in the first line of Figure 36, then there will be a read in `p2.fastq` with the identifier `@HWI-EAS348:12:FC:1:1:2201:3496 1:N:1:` (or any other digit different from `0`). Normally, in addition to the identifier, `FASTQ` files with paired-end reads preserve the ordering of the reads so that the first read in `p1.fastq` is paired with the first read in p2.fastq, and so forth. In this case, the identifier provides an extra level of quality control to make sure the reads match.

- **Line 2**: The nucleotides called by the sequencer.
- **Line 3**: Starts with the '+' character, normally followed by the same identifier as in Line 1. Because the identifier in Line 3 is optional, some FASTQ files only contain the '+'.
- **Line 4**: Indicates the quality scores of the nucleotides in Line 2 (they have a one-to-one correspondence). Each quality score is represented by an ASCII character and their values depend on the sequencer machine. The following section discusses the quality scores in detail.

One final note, in Figure 36, we see marked in blue the tail that corresponds to a forward primer (T7) and in red the reverse complement of the tail of a reverse primer (T3c). These are the tails attached to the primers illustrated in Figure 35.

5.4.2 Quality scores assigned to reads

Each nucleotide in a read has a quality score associated to it. This score, known as Phred score, is a measure of the *certainty* with which the sequencer made the call for a particular base. Although the newer Illumina sequencers can assign Phred scores between 2 and 62, FASTQ files currently have a maximum score of 40. In order to represent these scores as a printable ASCII character, the scores are offset by 33. For example, a quality score designated by the character # in the second read of Figure 36 corresponds to a Phred score of 2 = 35 - 33 = ASCII_code(#) - 33. Similarly, the quality score assigned to the character H is 39.

The relationship between a Phred quality score Q_{Phred} and the probability p that the base call is incorrect is modeled by the following equations (Cock et al., 2010):

$$Q_{Phred} = -10log_{10}(p)$$

or equivalently for p

$$p = 10^{-\frac{Q_{Phred}}{10}}$$

Therefore, following the previous example, for a value of $Q_{Phred} = 2$ –in the low end of quality scores– we have $p = 0.316$. That is, there is a chance of almost 1 in 3 that the base call is incorrect. Likewise, for $Q_{Phred} = 39$ –in the high end of quality scores– we have $p = 0.0001$ (a much smaller chance of 1 in 10,000 that the call is incorrect).

The quality scores are indispensable when mapping the reads to the reference genome. Almost all mapping algorithms allow for minor discrepancies (insertions/deletions) between the read and the reference genome. For example, if the mapper is unable to map exactly the sequence `AACTGGA`, it will try to map variations of it with possible insertions (e.g.: `AAC·TGGA`) or deletions (e.g.: $\texttt{AA}\cancel{\texttt{C}}\texttt{TGGA}$) These discrepancies are weighted based on the quality score of the nucleotides. Nucleotides with very high quality scores are expected to match exactly when mapped, whereas nucleotides with low scores are assumed to arise from an incorrect call and are seldom matched. As a result of this, reads containing a large number of nucleotides with low quality are normally discarded by the mapper.

5.4.3 Mappability of restriction fragments in Igh

As mentioned earlier, a key component in the detection of DNA interactions using 5C is the sequencing of reads and their posterior mapping to a reference genome. If the quality of sequencing is poor or if the paired-end reads cannot be mapped uniquely to the genome, then the interactions captured through the chemical process will not be reported.

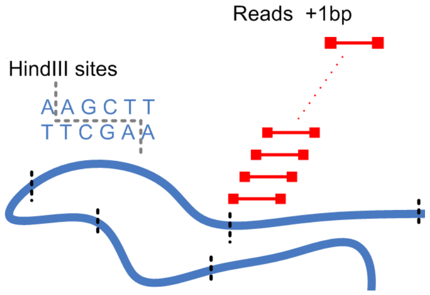
We decided to analyze all the restriction fragments in the mouse genome to determine their *mappability*. For each fragment, we created artificial reads of lengths 75, 100 and 150 bp. The reads spanned the entire fragment with a sliding window of 1 bp, depicted in Figure 37(a). We then attempted to map the reads to the mouse genome (mm9) using `Bowtie` as mapper (Langmead et al., 2009). Reads that mapped to more than one location within the fragment or to different fragments were discarded. The mappability index of each fragment was defined by the formula in Figure 37(b):

We computed the mappability index for the entire Igh locus in a similar way as described above. Table XXIII contains the results.

TABLE XXIII

MAPPABILITY OF RESTRICTION FRAGMENTS IN THE IGH LOCUS.

	Before alignment	Aligned		Failed	
Read length	# Reads	# Reads	%	# Reads	%
75	2,685,667	2,337,776	87.0%	347,891	13.0%
100	2,667,428	2,426,730	91.0%	240,698	9.0%
150	2,631,233	2,489,554	94.6%	141,679	5.4%



$$mappability(frag_i) = \frac{\#\text{ mapped reads to} frag_i}{\#\text{ reads from} frag_i}.$$

(a) Generation of virtual reads (step +1 bp)

(b) Mappability score of restriction fragment i

Figure 37. Computation of mappability scores for restriction fragments.

Unfortunately, for reads of length = 75 bp, it would be impossible to recover all interactions in the locus. This is due to the fact that this locus contains regions with large number of repeats. DNA interactions on which these regions are involved will not be successfully detected if we attempt to map the reads directly to the genome. On the other hand, and as expected, when we consider reads of longer length the mappability improves. A read of 150 bp can be generated by the new Illumina HiSeq 2500/1500 sequencer, although currently the majority of the data available were generated with the more common length of 75 bp.

The above mentioned mappability scores give us an idea of the best case scenario in which a read, with perfect quality, is mapped back to the genome. In our particular case, after removing the forward and reverse tails, we only have reads of 22 bp to map. Additionally, instead of looking at the entire restriction fragment, we should focus on the fragment ends where the primers are located. If for reads of 75 bp we had an overall mappability of 87%, for 22 bp we will surely have a much lower mappability. Our intuition is confirmed in Section 5.4.5 where we perform a first attempt at mapping our 5C data to the entire mouse genome.

5.4.4 Trimmed FASTQ reads

Before mapping the paired-end reads to the mouse genome we need to remove the forward and reverse tails. This process is known as **trimming** and consists of altering the FASTQ files by removing nucleotides from Line 2 and their matching quality scores in Line 4. A trimmed version of the FASTQ file in Figure 36 is depicted in Figure 38.

```
@HWI-EAS348:12:FC:1:1:2201:3496 1:N:0:
ATAAATTACTGAAGGGCTTAAG
+
H@HHGHH<GHBGG@GFEHHHHD
@HWI-EAS348:12:FC:1:1:2201:11118 1:N:0:
AGCACAGCCTGATAAACACCTT
+
G8GG2-;/843;.>9EE#####
```

Figure 38. Trimmed version of the reads in Figure 36 (forward/reverse tails were removed) .

In general, trimming is performed to remove a barcode of length k or other chemically added sequence of DNA nucleotides (e.g.: the forward and reverse tails in our 5C data). In this case, k nucleotides are removed from all sequences regardless of their quality. Another common approach to trimming consists in removing from each read a variable number of nucleotides depending on their quality. Because of technical characteristics of the sequencing technology, the 5'end of a read (left portion) will have better quality scores than the 3'end (Shendure and Ji, 2008; Mardis, 2008). Variable trimming will, therefore, remove more nucleotides from the 3'end (right portion).

Going back to our 5C data, we had to assess if the quality of our reads warranted the use of variable trimming, especially in the 3'end of reads. To that effect, we analyzed the original `FASTQ` files and quantified the Q_{Phred} scores at each position: from 1 to 42. We created, per replicate, boxplots of these scores and Figure 39 shows the results for pro-B replicate 1. Refer to Figure 66 in the Appendix for results in MEF.

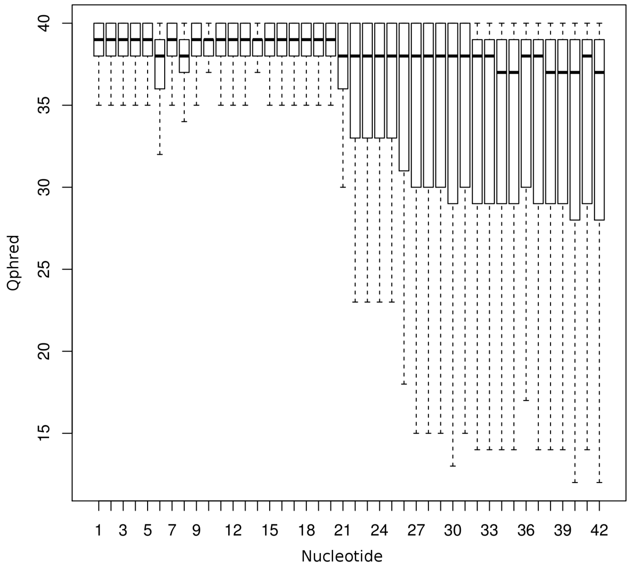


Figure 39. Quality scores of nucleotides (length of read = 42) in reads of pro-B replicate 1.

Although, as expected, there is a decay in quality towards the 3'end, the median Q_{Phred} at nucleotide 42 is 37 (mean = 29.8). In light of these high quality scores, we did not implement quality trimming. We simply trimmed 20 nucleotides from the 5'end to remove the forward and reverse tails as depicted in Figure 38. These trimmed dataset constitutes the reads that will be mapped to the genome. The question we address in the next two sections is: What is *the best reference genome* against which the reads should be mapped?

5.4.5 Mapping trimmed reads to the mouse genome

The trimmed `FASTQ` files were then mapped to the mouse genome (mm9) using `Bowtie`. Each paired-end (PE) was mapped independently of the other. Reads that mapped to more than one location were discarded as they would create ambiguous interpretations. Table XXIV shows the results from the mapping.

TABLE XXIV

MAPPING READS TO THE MOUSE GENOME (MM9).

		Number of reads				
Cell	Replicate	Original	Mapped PE1	%	Mapped PE2	%
pro-B	1	25,175,204	5,966,854	24%	5,874,280	23%
	2	25,919,640	5,753,016	22%	5,857,633	23%
MEF	1	26,537,062	7,014,551	25%	7,098,591	27%
	2	24,766,329	6,728,370	27%	6,953,853	28%

We can therefore conclude that our mapping was very ineffective ($\sim$ 22-28% of successfully mapped reads). As we anticipated from the analysis of mappability scores in Section 5.4.3, reads that are mapped directly to the Igh locus would yield poor results. If you recall, in that section we tested the smallest read length of 75 bp assuming perfect quality scores. Here we are attempting to map reads of 22 bp of length obtained from the 3'end of the read, i.e., with lower quality scores. It is clear from these results that we need a better mapping approach.

5.4.6 Mapping trimmed reads to a "virtual" genome

Instead of mapping the reads to the entire mouse genome, we can make use of our understanding of the 5C experiment and map the reads to a "virtual" genome created from the 5C primers. `Bowtie` provides a tool (`bowtie-build`) to construct indexes of any genome. The index is what `Bowtie` uses to efficiently map reads. We used the sequence information of all the 5C primers designed for our experiment and `bowtie-build` to construct a virtual genome of these primers. Reads mapped to the virtual genome could only map to a known primer, thus removing any spurious DNA sequences captured by 5C. Additionally, we required reads to map to only one primer, discarding any reads with multiple mappings.

Because we have the genomic coordinates of each 5C primer, after a read is successfully mapped to our newly created genome, we can tell exactly the genomic location to which the read mapped. Table XXV shows the results obtained from mapping to our virtual genome.

Each end of a paired-end read was mapped independently of the other and their results are shown in the columns "Mapped PE1" and "Mapped PE2" of Table XXV. Reads where both ends mapped successfully, went through an extra filtering stage that required one end to map to a forward primer and the other end to a reverse primer. Very few cases were found where both mapped ends mapped to two forward (or reverse) primers serving this as a validation of the correctness of the 5C experiment.

TABLE XXV

MAPPING STATISTICS.

		Number of reads						
Cell	Replicate	Original	Mapped PE1	%	Mapped PE2	%	Matched	%
pro-B	1	25,175,204	23,034,072	92%	22,593,323	90%	21,424,015	85%
	2	25,919,640	23,420,680	90%	23,024,139	89%	21,964,806	85%
MEF	1	26,537,062	24,939,514	94%	24,651,205	93%	23,858,581	90%
	2	24,766,329	23,358,067	94%	23,151,825	94%	22,332,216	90%

Due to its high mapping yield, this is the mapping strategy we adopted and the results of our analysis are based on this mapping methodology. After the reads are mapped to the virtual genome and filtered as mentioned above, the next step consists in matching the paired-end reads. If only one end of the paired-end can be successfully mapped, then the entire pair is discarded. Matching pairs are then quantified by creating a matrix $\boldsymbol{\mathcal{M}}_{primer} \in \mathbb{N}^{F \times R}$ where F and R are the numbers of forward and reverse primers respectively. If a matched pair is of the form (fwd_i, rev_j) where the first element corresponds to forward primer i and the second element is reverse primer j, we increment by one the count at $\boldsymbol{\mathcal{M}}_{primer}[i, j]$. After processing all matched pairs, $\boldsymbol{\mathcal{M}}_{primer}$ contains the total number of interactions between all forward and reverse primers.

Obtaining the matrix $\boldsymbol{\mathcal{M}}_{primer}$ is the end of our pre-processing stage and it is the input to all the downstream analysis steps detailed in Sections 5.7–5.10.

5.5 Example to illustrate visualization methods

Before delving into the details of the methods we have developed, we would like to take a brief detour to discuss different visualization techniques we implemented to interpret interactions in 5C data. We want to walk you through a toy example that summarizes the topics discussed in Section 5.4 and that will help you better understand the peculiarities of these data. By the end of this section we expect you to be more familiarized with the terminology and to have an intuition for the need of the algorithms we propose later on.

We start by defining a genomic area of interest in a chromosome. We want to capture DNA interactions that occur between the genomic coordinates 1 and 3000. Figure 40(a) depicts this region. The horizontal green line represents the chromosome and the ticks are the **restriction enzyme sites** (e.g. HindIII sites). Restriction enzymes are proteins that cut DNA at specific sites and the ticks in the figure symbolize these sites. For example, HindIII is a restriction enzyme that recognizes the pattern 5'-`AAGCTT`-3' and cleaves DNA between the first and second adenine. The segment of DNA in between two restriction sites is called a **restriction fragment**.

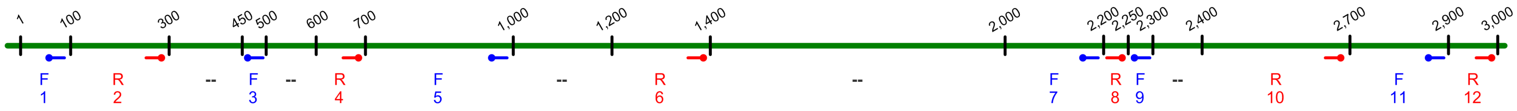
In Figure 35 we saw two restriction fragments interacting via two proteins. The interaction in the figure occurs (approximately) in the middle of the fragments. In reality, an interaction can take place at any position but because of limitations with the technology, we can only identify the interacting fragments through their primers (marked in blue and red in Figure 35). That is to say, the granularity of this method is at the restriction fragment level.

In addition to the restriction fragments, Figure 40(a) shows the location of **forward** (F) and **reverse** (R) **primers**. Their alternating pattern was designed to allow for the detection of a dense matrix of interactions throughout the genomic region of interest (Dostie et al., 2006).

Not all restriction fragments have a primer associated to them and this can mean: a) we are not interested in detecting interactions that involve those fragments or b) we cannot create a primer for the fragments due to the presence of repeats. In Figure 40(a), the fragments with no primer are labeled as “`--`”.

The two tables in Figure 40(b) and Figure 40(c) provide details about the forward and reverse primers respectively. Moreover, they also indicate the restriction fragments the primers recognize. We adopted the convention of starting a fragment +1 bp from the preceding restriction enzyme site.

We assume an intermediate process, not shown in the figure, that reads a `FASTQ` file containing paired-end reads, maps them to a virtual genome created with the DNA sequences of the primers, and quantifies all the interactions detected between any pair of forward-reverse primers. The output of this process is the matrix $\boldsymbol{\mathcal{M}}_{primer}$ which we described in Section 5.4.6. This matrix has as many rows and columns as there are forward and reverse primers respectively. Following our example, Figure 40(c) shows $\boldsymbol{\mathcal{M}}_{primer}$ with the number of interactions we assume to have detected between the primers.



(a) Alternating forward (blue) and reverse (red) primers. Not all restriction fragments have a primer.

Forward primers			
primer_id	**frag_id**	**start**	**end**
1	frag_401	2	100
3	frag_404	451	500
5	frag_407	701	1000
7	frag_411	2001	2200
9	frag_413	2251	2300
11	frag_416	2701	2900

(b) Coordinates of forward primers

Reverse primers			
primer_id	**frag_id**	**start**	**end**
2	frag_402	101	300
4	frag_406	601	700
6	frag_409	1201	1400
8	frag_412	2201	2250
10	frag_415	2401	2700
12	frag_417	2901	3000

(c) Coordinates of reverse primers

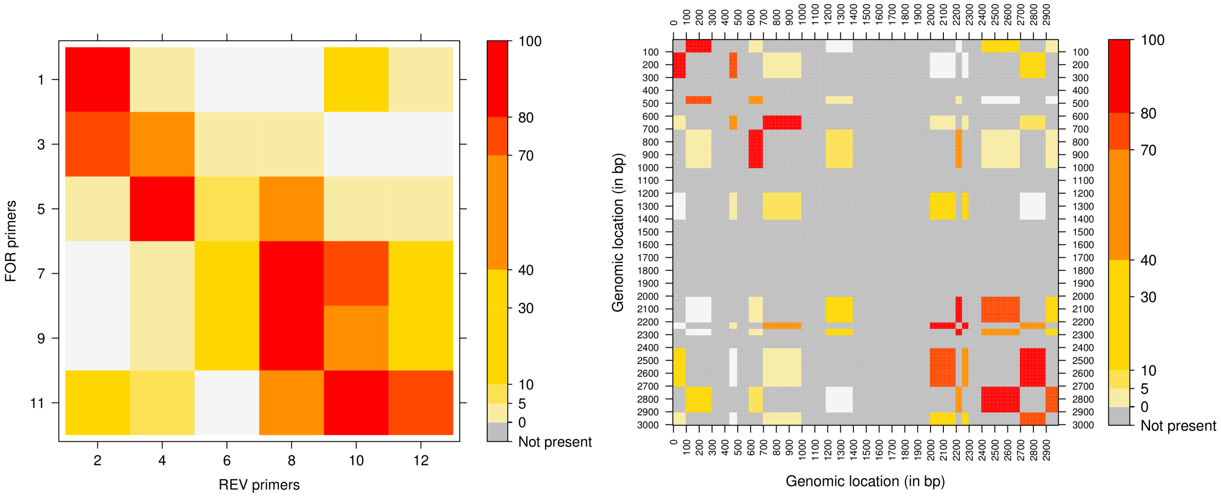
		Reverse primers					
		2	4	6	8	10	12
	1	100	2	0	0	30	3
	3	80	70	5	2	0	0
Forward primers	5	5	100	10	60	5	1
	7	0	3	40	100	80	30
	9	0	1	15	100	60	20
	11	40	8	0	50	100	80

(d) Matrix with interactions between primers: $\boldsymbol{\mathcal{M}}_{primer}$

Figure 40. Sample 5C data to illustrate visualization methods.

5.5.1 Visualizing raw contact maps

A first attempt at visualizing these data is shown in Figure 41(a), which is basically a heatmap of the matrix in Figure 40(c). It is called a **primer contact map** because it shows the interactions between the primers without consideration of their spatial location in the genome, only their relative ordering. The Y-axis represents the forward primers, with one tick per primer. Similarly, the X-axis shows the reverse primers. The topmost left corner of the heatmap is its origin, i.e.: interaction of the first forward and first reverse primers.



(a) Primer contact map (no genomic scale, asymmetric)

(b) Fragment contact map (with genomic locations, symmetric)

Figure 41. Raw contact maps used to illustrate visualization methods (data from matrix in Figure 40(c).

An immediate improvement to the primer contact map is the **fragment contact map** in Figure 43(b). This heatmap shows the same interactions from the primer contact map but with respect to the restriction fragments where the primers fall. Additionally, because the X and Y axis represent genomic coordinates, the fragment contact map shows true interactions between genomic regions. For example, from the matrix in Figure 40(c), forward primer 1 (F1) interacts with reverse primer 10 (R10) a total of 30 times. F1 falls in fragment frag_401 that spans from location 2 to 100 (Figure 40(b)). Similarly, R10 is in fragment frag_415 which goes from location 2,401 to 2,700. With the origin of the heatmap in the topmost left corner, the interaction between these two fragments can be found in the X-axis between 2 and 100 and in the Y-axis between 2,401 and 3,000 (or vice versa since the heatmap is symmetric).

The color grey in the figure means there are no primers in those locations and, therefore, no interactions can be detected. This is referenced as "Not present" in the key to the right of the heatmap. The most important feature of the fragment contact map is its ability to show these regions without primers. It provides a quick view of the density of primers and their interactions in different regions. In the bottommost right corner of the map, from location 2,000 to 3,000 we can see a larger density of primers, reflected by the fragments with interactions in that area. On the other hand, the center of our heatmap (from 1,400 to 2,000) is grey due to a lack of primers.

5.5.2 Visualizing binned contact maps

The most salient feature of the fragment contact map can also be one of its caveats. If our genomic region of interest is large, say in the order of several megabases and our density of primers is low, the fragment contact map will be almost entirely grey. There will be a few scattered colored points indicating the fragments where the primers fall but the interactions will be hard to visualize and the map will not convey useful information. The solution to this problem is to bin the data creating **binned contact maps**. Two parameters are needed to create a binned contact map: a) the bin size and b) the step size.

Binning consists in applying two sliding windows of the same size. One window will be referred to as the "Forward window" and it is used to recognize the presence of forward primers. The other will be the "Reverse window" and it will recognize reverse primers. Let's assume we want to bin our data using a bin size of 300 bp (window size) and a bin step of 10 bp. Before we begin to slide the windows, we need to create an empty square matrix $\boldsymbol{\mathcal{M}}_{binned} \in \mathbb{R}^{n \times n}$ where $n = \lceil \frac{length\_region}{bin\_step} \rceil$. Following our example from Figure 40, we have $length\_region = 3,000$ and we are assuming $bin\_step = 10$, so our matrix $\boldsymbol{\mathcal{M}}_{binned}$ will have 100 rows and columns. $\boldsymbol{\mathcal{M}}_{binned}$ is defined in the domain of real numbers because, as it will become clear in the next paragraph, binning can create non-integer counts for the interactions.

The process of binning begins when the center of both windows is at the beginning of our genomic region of interest, i.e.: location 1 bp. For the moment, ignore the fact that half of the window is outside our region. Any forward primers, in the region overlapping with the Forward window, are retained. Likewise, reverse primers that overlap with the Reverse window are kept. If we only have one forward and one reverse primer, then we simply obtain the count of interactions from the matrix $\boldsymbol{\mathcal{M}}_{primer}$ (see sample matrix in Figure 40(d)). This value is then stored in the matrix $\boldsymbol{\mathcal{M}}_{binned}$, in the row and column corresponding to the position of the bins (row=1 and column=1 when the program starts). If we have more than one forward and/or reverse primers overlapping with the windows, we obtain the individual counts of interactions between each forward-reverse primer pair and we compute the mean. Then, as before, this value is stored in $\boldsymbol{\mathcal{M}}_{binned}$. It is this step than can create non-integer counts of interactions between two bins.

In the next iteration we move the Reverse window *bin_step* base pairs to the right and repeat the process. Once the Reverse window has been slid to the end of the region, we reset its position to the beginning and we slide the Forward window *bin_step* base pairs to the right. The binning process ends when the Forward window reaches the end of the region.

$\boldsymbol{\mathcal{M}}_{binned}[i, j]$ will contain the number of interactions found at the intersection of bin i and bin j. Because the 5C interactions have no directionality, if we have c interactions between bins i and j, then we also have c interactions between bins j and i. Figure 42 illustrates the methodology of binning.

The reason why the windows have to be centered at each of the bins is because the algorithm will otherwise "shift interactions" to the left. An example of the binned heatmap we were discussing before can be found in Figure 43(a). In order to create more contrast between the colors, the color assigned to the largest range of interactions was changed from red in Figure 41 to dark red in Figure 43.

It is interesting to compare Figure 41(b), which showed the restriction fragments, to Figure 43(a), which shows the binned interactions. Due to the fact that the sliding windows grab interactions from neighboring sites, the amount of grey in binned contact maps is less than in

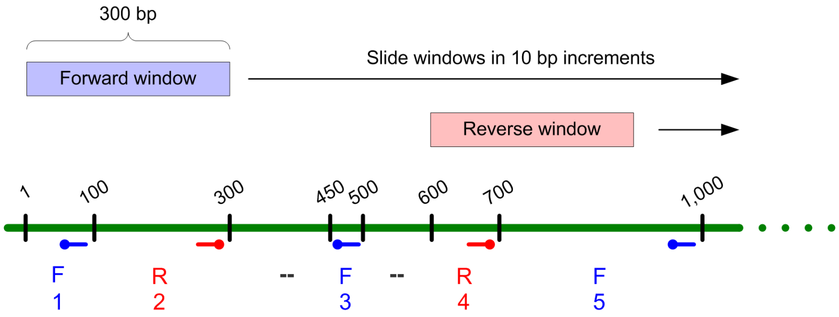
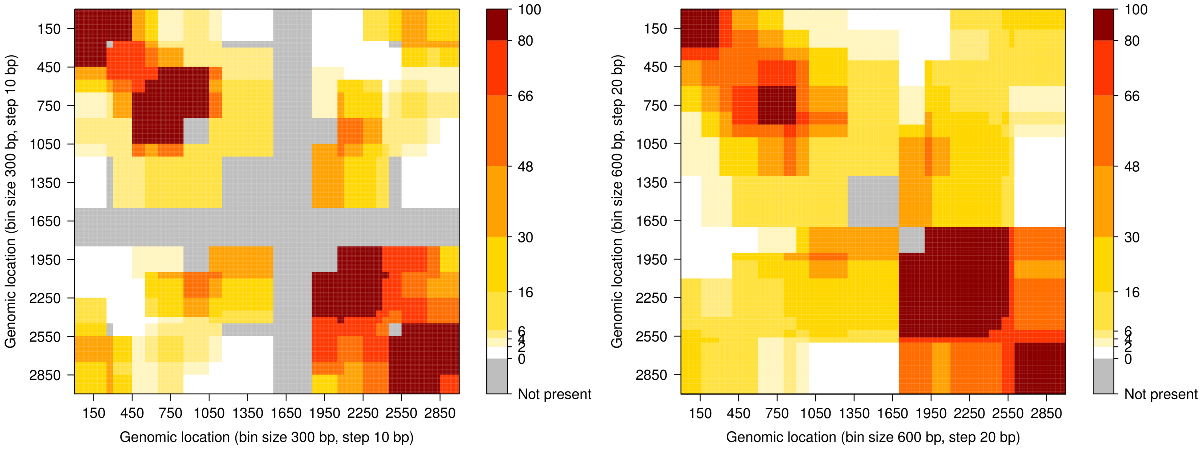


Figure 42. Example of a sliding window when creating a binned contact map (bin size=300, bin step=20).

fragment contact maps. This of course, has some pros and cons. On the bright side, when our region of interest is large, a binned contact map is easier to understand than a fragment contact map because it accentuates the areas where the larger number of interactions occur. Conversely, and this is its main drawback, a binned contact map “predicts” interactions where none exist. The binning step is equivalent to smoothing the data and this the reason why there is less grey in Figure 43(a).

If we repeat the binning using a wider window ($bin\_size = 600$) and longer step ($bin\_step = 20$), we will obtain the binned contact map shown in Figure 43(b). This time, only the center region appears to lack any interaction information and the more demarked interacting regions found in Figure 43(a) become more spread out.

Because of its advantages in helping to visualize the data, and despite its shortcomings at creating “fictitious” interactions, the binned contact maps have become a de facto standard in



(a) Bin size=300 bp, bin step=10 bp

(b) Bin size=600 bp, bin step=10 bp

Figure 43. Binned contact maps illustratrating visualization of interactions (data from matrix in Figure 40(c).

visualizing long range chromatin interactions (Lajoie et al., 2009; Wang et al., 2011; Sanyal et al., 2012).

5.6 Primer design in Igh locus

Now we return to our 5C experiment to explore the coverage of the primers in the Igh locus and in the gene desert of chr5. The location of the primers for both loci were defined in Dr. Kenter's lab. The primers follow an alternating pattern (Dostie et al., 2006) as illustrated in Figure 40(a) and were designed using the web tool My5C (Lajoie et al., 2009).

The Igh locus contains 112 and 113 forward and reverse primers respectively. Because primers capture the interactions of the restriction fragments to which they belong, we obtained the minimum and maximum genomic coordinates of the restriction fragments covered by the primers. Table XXVI summarizes this information for Igh and the gene desert in chr5.

TABLE XXVI

NUMBER OF 5C PRIMERS AND CHARACTERISTICS OF THE RESTRICTION FRAGMENTS THEY COVER.

	Primer		Coordinates of fragments covered		
Locus	Type	Count	Min.	Max.	Span
Igh (chr12)	Forward	112	114,341,250	117,346,617	∼ 3 Mb.
	Reverse	113	114,339,371	117,349,198	
Gene desert (chr5)	Forward	24	133,181,426	133,463,499	∼ 290 Kb.
	Reverse	25	133,174,212	133,464,772	

The area of the Igh locus on which we will focus is 3 megabases wide. Of the total of 225 restriction fragments covered by either a forward or reverse primer, we obtained some statistics about their lengths and of the gaps between them. With respect to the fragments' length Table XXVII summarizes these findings and Figure 44 shows their distribution. Refer to the Appendix, Figure 67 for a histogram with the lengths of fragments in chr5.

TABLE XXVII

CHARACTERISTICS OF RESTRICTION FRAGMENTS IN OUR 5C EXPERIMENT.

			Frag. length				Frag. length	
		Frag.	(< 10 Kb)				(> 10 Kb)	
Locus	Primer	count	Median	μ	σ	Max.	Count	Max.
Igh (chr12)	Forward	112	2,688	3,386	2,259	9,658	6	20,182
	Reverse	113	2,731	3,318	2,068	9,827	6	13,858
Gene desert (chr5)	Forward	24	4,303	4,542	2,588	9,982	1	12,971
	Reverse	25	3,567	4,134	2,830	9,939	2	12,952

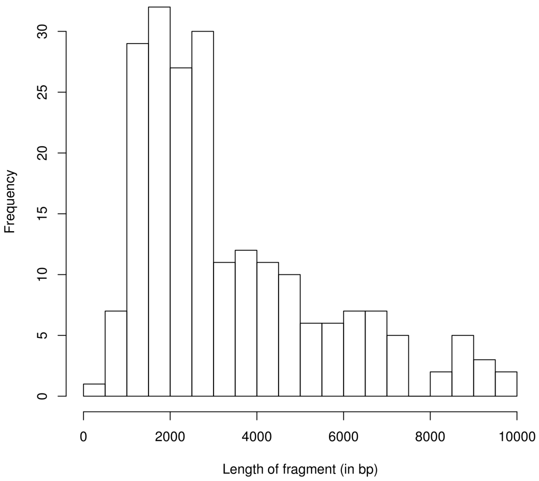


Figure 44. Length of restriction fragments covered by a primer (forward or reverse) in the Igh locus (chr12). Bin size=500 bp.

The length of the fragments is of particular importance as it may have an effect on the efficiency of ligation (the third step described in Figure 35). In fact, this is one of several experimental biases that have been found to occur in Hi-C experiments (Yaffe and Tanay, 2011) and in 5C experiments (Sanyal et al., 2012). Figure 45 illustrates this case.

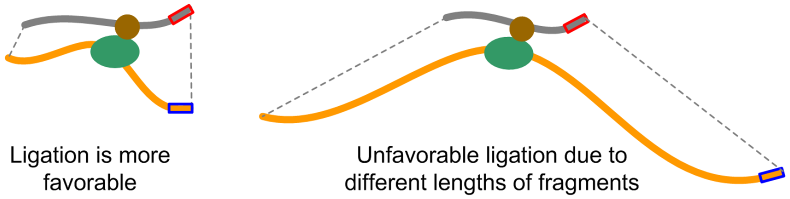

Ligation is more favorable
Unfavorable ligation due to different lengths of fragments

Figure 45. Possible 5C experimental bias with respect to length of fragments during ligation.

The ligation of two restriction fragments whose lengths are similar is more favorable than when one fragment is much longer than the other. In our case, the minimum length of a fragment with a forward primer is 305 bp whereas the maximum fragment length with a reverse primer is 13,858 bp. Similarly, the minimum fragment length (reverse) is 601 bp while the maximum fragment length (forward) is 20,182 bp. Nevertheless, because of the conditions on which the 5C experiment was conducted and the experimental protocol observed, we do not believe the differences in fragment length have the potential of masking true interactions. As a result of it, this potential source of bias was not considered in posterior analysis steps.

In addition to analyzing the length of fragments, an interesting metric is the length of the gaps between the fragments. This can be used as a measure of how sparsely located the primers are. To determine the gaps, we grouped together all primers in a locus and obtained statistics of the gaps between a forward and reverse primer. Because of their alternating nature, the gaps represent the distances between any two neighbors. Table XXVIII lists the gap statistics.

TABLE XXVIII

CHARACTERISTICS OF THE GAPS BETWEEN RESTRICTION FRAGMENTS IN OUR 5C EXPERIMENT.

		Gap length				
Locus	Gap count	Median	μ	σ	Min.	Max.
Igh (chr12)	224	1,005	9,583	18,530	0	142,700
Gene desert (chr5)	48	0	1,140	3,308	0	16,046

From the table, it seems that the distribution of primers in the gene desert is more compact than in Igh. The median gap length in chr5 is 0 which means the fragments are contiguous. In fact, 31 out of a total of 48 fragments with primers in chr5 are contiguous (65%). This should be contrasted to Igh with only 90 out of 224 contiguous fragments (40%). The histograms with the distributions of gaps in Igh and chr5 can be found in the Appendix (Figure 68 and Figure 69 respectively)

To conclude our analysis of restriction fragments, we obtained an objective measure of coverage in our genomic areas of interest. We computed a **primer density** by counting the number of nucleotides in each restriction fragment containing a primer and normalizing with the total number of nucleotides in the region. The normalized values were then binned with a bin size of 100,000 bp and a density curve was obtained using Gaussian smoothing. Figure 46 shows the primer density in the Igh locus.

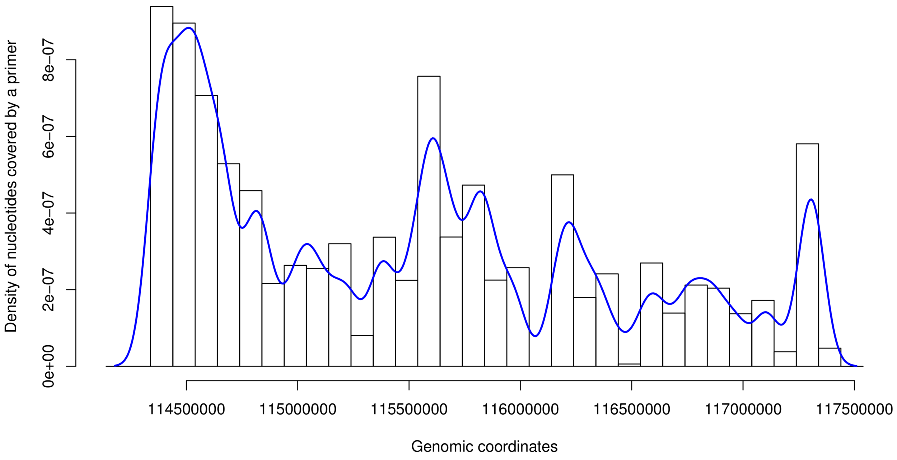


Figure 46. 5C primer density in the Igh locus (chr12).

It is clear from the picture that the beginning of the region is more densely packed with primers than other regions in the locus. This is an important fact because not only will we be able to detect more interactions in this area but also it will have an effect on our calculations of the overall *expectated number of interactions*. After a peak at the beginning of the Igh locus, we see a decay in the density of coverage towards the end of the locus. Note the region with almost no primers around 116,500,000 and recall from Table XXVIII the maximum gap length of 142,700 bp.

In contrast to Igh, the gene desert in chr5 seems to maintain the same density of coverage across its 290 Kb. This can be seen in Figure 47 (bin size=10,000 bp).

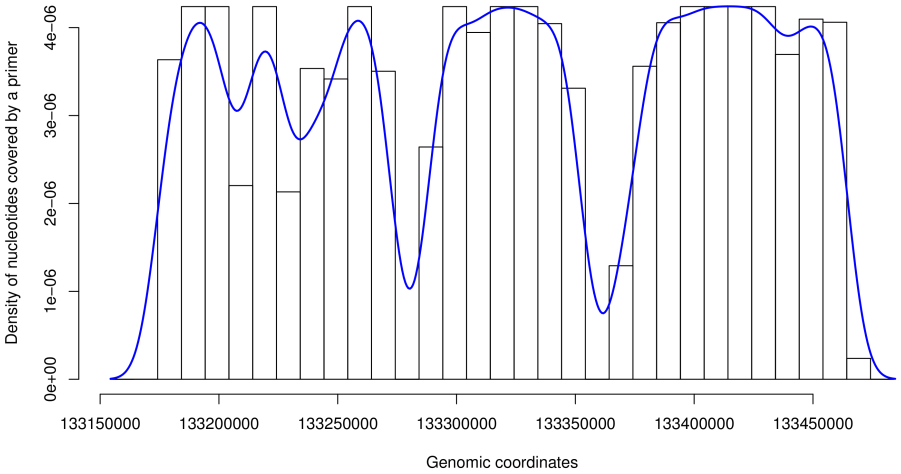


Figure 47. 5C primer density in the gene desert of chr5.

The fact that the gene desert is densely packed with primers is important and it leads us to the next section, where we develop a methodology to use the interactions detected in the gene desert to scale our 5C data.

5.7 Using an internal control to normalize interactions in 5C

To recapitulate where we stand, from a processing point of view:

1. We have 5C data from 4 independent biological samples:
   - 2 samples of pro-B cells
   - 2 samples of MEF cells.
2. In each sample, interactions are probed in two different regions:
   - Igh locus: our region of interest
   - Gene desert in chr5: used as internal control
3. **Goal**: determine what interactions between primers i and j in Igh are different between pro-B and MEF.

Keeping in mind that our ultimate goal is to compare interaction frequencies in different cells, we must account for potentially random experimental variations that arise from the use of the technology. Some of these experimental variations are (Fraser et al., 2012):

- Difference in cutting efficiency of the restriction enzymes: even for the same cell type, minor conformation changes can render chromatin regions less accessible to the restriction enzyme and this will affect their overall efficiency.
- Amount of 3C template used to generate the 5C library.
- Difference in total number of reads obtained from the sequencer.

We are faced with the problem of comparing a signal in different biological samples, when the true signal may be obscured by noise introduced from technical problems in the use of the technology. Drawing a parallel with a different technology, issues like this one were detected more than a decade ago at the early stages of microarray analysis. For example, when using two-color microarrays, normalization was used to minimize biases associated to differences in the fluorescence of dyes. Normalization methods were classified as (Yang et al., 2002): a) **global normalization** when the two signals in the array –red (R) and green (G)– were corrected with a constant factor so that $R = kG$ where k was a globally obtained median of log-intensity; and b) **local normalization** that corrected signals based on local spot intensity using linear methods such as simple regression or non-linear methods such as LOWESS (Cleveland, 1979). Similarly, the method of scale-normalization (Smyth and Speed, 2003) was a between-array-normalization procedure to scale log-ratios of expression from a group of two-color cDNA arrays. This scaling guaranteed that each array had the same median absolute deviation. In essence, all of these methods aimed at putting the microarrays on the same playing field in order to compare gene expression across samples.

Despite the fact that we are dealing with a different technology, our research question is not different from the question the microarray methods tackled. We have a count of $I_{i,j}^{proB}$ interactions between primers i and j in pro-B and $I_{i,j}^{MEF}$ interactions between the same primers in MEF. How can we scale these counts in order to be able to compare directly $I_{i,j}^{proB}$ and $I_{i,j}^{MEF}$?

Not surprisingly, some methods that have been suggested to normalize 5C data, borrowed on the ideas and methods developed for microarrays. A variation of global normalization is used to normalize differences between two cell types in 3C experiments (Dekker, 2006). The method suggests to use an internal control with a set of interactions assumed to be the same in both cells e.g.: a housekeeping gene. A similar approach recommends to use **gene deserts** as internal controls (Fraser et al., 2012) and assumes that the chromatin architecture of a gene desert will not have significant changes between different cells. Both methods (Dekker, 2006; Fraser et al., 2012) agree on using the internal control to derive a scaling factor which will then be used to normalize the region of interest.

Other normalization methods take a different approach, for example, by computing a Z-score between interacting primers i and j (Baù et al., 2011). The score is defined as Z-score$_{ij} = \frac{\mu - f_{ij}}{\sigma}$ where $f_{ij} = log_{10}$(interactions between fragments i and j); μ and σ are the mean and standard deviation of log_{10} frequencies in the interaction matrix. A different normalization technique corrects interactions between primers i and j by computing two scaling coefficients c_i^{intra} and c_j^{intra} based on the intra-chromosomal interaction efficiency of each of them individually (Sanyal et al., 2012). Once these coefficients have been computed, the corrected frequency of interactions between i and j is computed as $\hat{f}_{ij} = f_{ij} \cdot c_i^{intra} \cdot c_j^{intra}$. It is important to note that these last two methods do not rely on an internal control.

The normalization procedure we propose is based on the use of a gene desert to scale the data in the Igh locus, similarly to the first two methods described above (Dekker, 2006; Fraser et al., 2012). If you recall, their key feature was to use an internal control to scale the data.

Let us illustrate with an example how this idea of scaling works as it is at the core of the normalization procedure propose by us.

5.7.1 Example of derivation of a scaling factor using a gene desert

We assume we have two different cell types: $\boldsymbol{B}$ and $\boldsymbol{M}$. Both cells share the same primer design and have F forward primers and R reverse primers defined in a gene desert. We will then have $P = F \cdot R$ possible pairs of forward-reverse primers. We can enumerate all possible pairs of primers $p = 1, 2, \ldots, P$ and we define $\boldsymbol{B}_p$ and $\boldsymbol{M}_p$ as the number of interactions for the pair of primers p in cells $\boldsymbol{B}$ and $\boldsymbol{M}$ respectively. For each primer pair p, we compute a log_{10}-ratio n_p between the two cells as in (Dekker, 2006):

$$n_p = log_{10}(\frac{\boldsymbol{B}_p}{\boldsymbol{M}_p}) \text{ with } p = 1, 2, \ldots, P$$

and then an average of all ratios:

$$n_{avg} = \frac{1}{P} \sum_{p=1}^{P} n_p$$

Finally, the scaling factor between $\boldsymbol{B}$ and $\boldsymbol{M}$ is defined as:

$$n = 10^{n_{avg}}$$

With numbers, let us assume we have 3 pairs of primers. Each pair of primers has the following number of interactions per cell type:

$$\begin{array}{lll} \boldsymbol{B}_1 = 13 & \boldsymbol{B}_2 = 4 & \boldsymbol{B}_3 = 15 \\ \boldsymbol{M}_1 = 21 & \boldsymbol{M}_2 = 7 & \boldsymbol{M}_3 = 32 \end{array}$$

Then, we compute the scaling coefficient $n = 0.5493837$

Finally, we can correct the data in either of the following ways:

$$\boldsymbol{M}_p \cdot n$$

$$\begin{array}{lll} \boldsymbol{B}_1 = 13 & \boldsymbol{B}_2 = 4 & \boldsymbol{B}_3 = 15 \\ \boldsymbol{M}_1 = 11.537058 & \boldsymbol{M}_2 = 3.845686 & \boldsymbol{M}_3 = 17.580278 \end{array}$$

$$\boldsymbol{B}_p \cdot \frac{1}{n}$$

$$\begin{array}{lll} \boldsymbol{B}_1 = 23.662879 & \boldsymbol{B}_2 = 7.280886 & \boldsymbol{B}_3 = 27.303322 \\ \boldsymbol{M}_1 = 21 & \boldsymbol{M}_2 = 7 & \boldsymbol{M}_3 = 32 \end{array}$$

Ideally, both $\boldsymbol{B}$ and $\boldsymbol{M}$ will have the exact number of interactions in the gene desert. This is because we are under the assumption that the gene desert acts as an internal control, maintaining the same conformation in different cells. But because of the experimental biases we mentioned at the beginning of Section 5.7, the number of interactions between the same pair of primers, in different cells, may differ. In the example above, there seems to be a 2-fold difference in the number of interactions between $\boldsymbol{M}$ and $\boldsymbol{B}$. The coefficient $n = 0.5493837$ is used to correct this difference.

A final step consists in applying the coefficient n to normalize interactions outside the gene desert. When the coefficient is used in this context, we will effectively be correcting for biases in the data. This correction will allow us to compare interactions between cell types which, in turn, was our original goal.

For our project, we implemented a similar mechanism to normalize 5C data in the Igh locus using the gene desert in chr5. In the next section we first describe the characteristics of the gene desert used in our experiment and then we proceed to describe our proposed normalization method.

5.7.2 Gene desert in chromosome 5

The genomic region in chromosome 5 that spans from 133,179,972 to 133,464,774 is considered a gene desert and was used as an internal control. This region, which was covered by 24 forward and 25 reverse primers has neither genes, microRNAs nor ORF mRNAs. Table XXIX summarizes all genomic elements that were checked for existence in the gene desert.

Because no known genomic elements were found in the region it can therefore be considered a true gene desert. The predicted elements are listed in Table XXX and were not considered relevant since there is no experimental validation about their existence or functional activity.

TABLE XXIX

GENE DESERT CHR5:133,179,972-133,464,774

Genomic Element	Name	Description
RefSeq genes	None found	
Non-mouse RefSeq genes	None found	
Mammalian Gene Collection Full ORF mRNAs	None found	
Vega Protein Genes	None found	
Vega Pseudogenes	None found	
microRNAs from miRBase	None found	
TROMER Transcriptome Database	None found	
Intl. Knockout Gene Consortium Gene	None found	
Ensembl Gene Predictions	ENSMUST00000157498	Predicted: RFAM and miRBase
tRNA Genes	U101853-1 U326585-1 U316042-1 U155315-1 U346204-1 U302262-1 U340601-1	BLAT alignment (mm9)

TABLE XXX

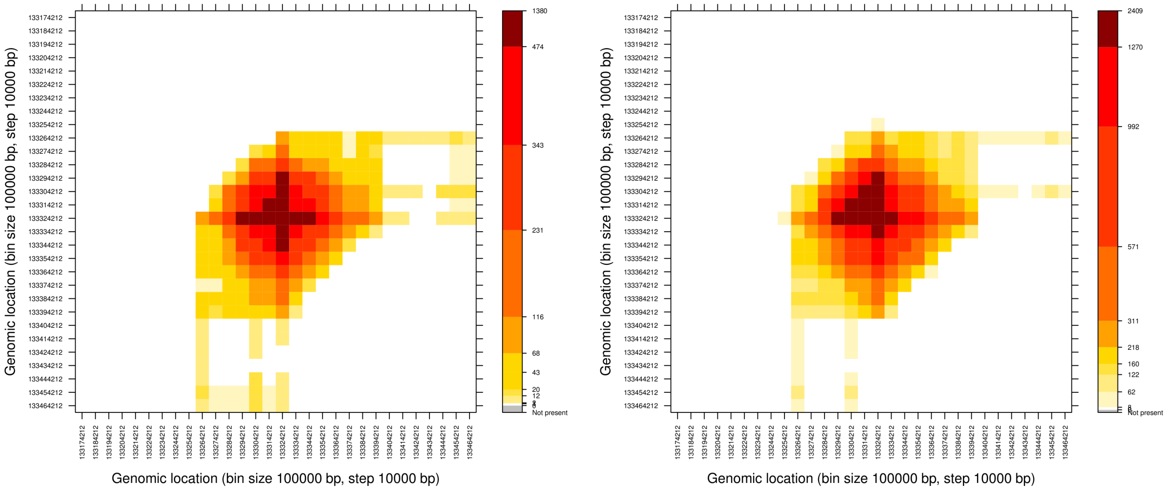
DETAILS OF PREDICTED ELEMENTS

Name	Definition	Length (in bp)
ENSMUST00000157498	Non-coding RNA member of clan 7SK found in metazoans	279
U101853-1	Stratagene mouse skin (#937313) cDNA clone	386
U326585-1	DAY10_16_K05.x1 FH DAY10	501
U316042-1	Stratagene mouse Tcell #937311	445
U155315-1	Stratagene mouse Tcell #937311	479
U346204-1	P2T1L5 Plasmodium yoelii infected liver tissues	134
U302262-1	AU259697 3'-directed mouse cDNA library	291
U340601-1	CJ309639 RIKEN full-length enriched mouse cDNA library	431

5.7.3 Binned contact map of the gene desert

Although there are no genomic elements in the gene desert, it may contain open chromatin which will fold, loop and will interact with itself and other genomic regions. These interactions will be captured by our 5C experiment. We want to determine the correctness of our assumption that the structure of the gene desert does not differ significantly between pro-B and MEF.

Figure 48 shows the interactions found in the gene desert of chr5 for pro-B and MEF (replicate 1). From visual inspection, the interactions in the cells seem to be very similar to each other. In the next sections we quantitatively analyze these interactions.



(a) pro-B, chr5.

(b) MEF, chr5.

Figure 48. Interactions captured in the gene desert of chr5 (replicate 1)

5.7.4 Number of interactions in the gene desert

Although the locations where the interactions take place are almost the same in pro-B and MEF, we can see from the keys next to the binned contact maps in Figure 48 that the magnitude of the interactions in MEF seems to be larger than in pro-B. We performed a systematic analysis by comparing the number of interactions between the same two primers i and j in pro-B versus MEF. We restricted our analysis to primers located within a distance greater than 12 Kb and less than 100 Kb from each other. The first condition (distance $>$ 12 Kb) is used to filter out **proximity events**. Portions of chromatin that are next to each other are more likely to interact in a random fashion and will not be good candidates to use as reference in our normalization process. The second condition (distance $<$ 100 Kb) is imposed to avoid long-range interactions which will be harder to consolidate.

The analysis was done on all possible combinations of replicates from pro-B and MEF. This is because although the samples are labeled replicate 1 and replicate 2, there is no relationship between replicate 1 in pro-B and, say, replicate 1 in MEF. The results for replicate 1 in both cells are shown in Figure 49. Refer to the Appendix, Figure 70, Figure 71, Figure 72 for details about the other combinations of replicates.

The left panel in Figure 49 shows the number of interactions between primers i and j in pro-B (X-axis) versus MEF (Y-axis). For each pair of primers we have a different count of interactions in pro-B and MEF. We analyzed the counts of the same pairs of primers between cell types and found a strong correlation between them: Spearman correlation coefficient $\rho = 0.781016$ with p-value=1.10e-19.

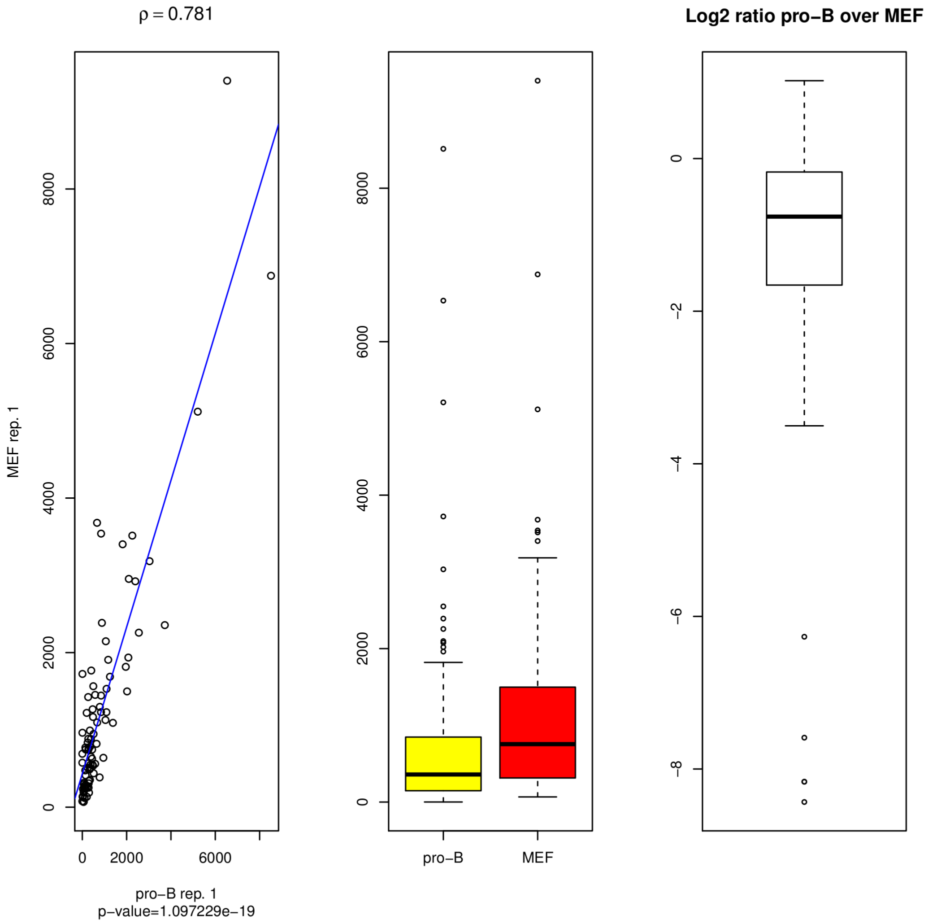


Figure 49. Interactions in chr5, replicate 1

The right panel also compares the counts of interactions between the same primers in pro-B and MEF. It is a boxplot of the log_2 ratios of pro-B over MEF. For example, if for primer pair $(i,\ j)$ we have 70 interactions in pro-B and 100 interactions in MEF, the plot will show $log_2(\frac{70}{100})$.

The center panel is a box plot of the count of interactions in each cell type (not paired). This, in addition to the skew towards negative values of the log2 ratios in the right panel, indicates that the number of interactions in MEF is larger than in pro-B.

Table XXXI shows the results of the correlation analysis for all combinations of replicates.

TABLE XXXI

CORRELATION ANALYSIS OF PRIMERS IN GENE DESERT (ALL REPLICATES).

Cell	Replicate	Cell	Replicate	Correlation coefficient	p-value
pro-B	1	MEF	1	0.781	1.10e-19
	2		1	0.772	5.02e-18
	1		2	0.923	2.50e-38
	2		2	0.880	1.42e-28

Based on the table, we conclude there is a very high level of agreement between primers in the gene desert (p-value $<$ 5.02e-18). This is very good news indeed, for in the next section we will use a subset of these primers to compute our normalization coefficients.

5.7.5 Method to obtain normalizing factors, using a gene desert with multiple replicates

In Section 5.7.1 we illustrated with an example an approach to normalize 5C data obtained from two different cells (Dekker, 2006). Our motivation to normalize the 5C interaction counts was to be able to compare them between cells. The example of Section 5.7.1 was fairly simple as we only had one replicate per cell. Here we propose a general method to obtain normalizing factors when the 5C experiment consists of multiple replicates.

Firstly, let us assume that for cell $\boldsymbol{B}$ we have $R_{\boldsymbol{B}}$ replicates, and for $\boldsymbol{M}$ we have $R_{\boldsymbol{M}}$ replicates. We define:

- $C = \{(x, y) | x \in \{1, 2, \ldots, R_{\boldsymbol{B}}\} \wedge y \in \{1, 2, \ldots, R_{\boldsymbol{M}}\}\}$ is the set of all possible combinations of replicates.
- an index set $K \subset \mathbb{N}$, such that $g : K \to C$ is a particular enumeration of the elements in C. This is equivalent to say that we have an indexed family $(c_k)_{k \in K}$.
- a function $f(c_k) = x$ where $x \in \{1, 2, \ldots, R_{\boldsymbol{B}}\}$ is the first element of the kth ordered-pair in C.
- a function $s(c_k) = y$ where $y \in \{1, 2, \ldots, R_{\boldsymbol{M}}\}$ is the second element of the kth ordered-pair in C.

Then, for a given pair p of primers (i, j) in the gene desert of chr5 we will have two interaction counts: $\boldsymbol{B}_p^{f(c_k)}$ and $\boldsymbol{M}_p^{s(c_k)}$, where $\boldsymbol{B}$ and $\boldsymbol{M}$ represent pro-B and MEF respectively; p is the primer pair; and c_k is a specific combination of replicates.

In our case, possible combinations are $C = \{(1, 1), (2, 2), (1, 2), (2, 1)\}$. As an example, if $k = 3$ then $c_k = (1, 2)$ and we will have $\boldsymbol{B}_p^{f((1,2))} = \boldsymbol{B}_p^1$ and $\boldsymbol{M}_p^{s((1,2))} = \boldsymbol{M}_p^2$ for pro-B replicate 1 and MEF replicate 2 respectively.

A **normalizing factor** for each combination of replicates can be computed as indicated in Equation 5.1 (Quackenbush, 2002).

$$n_k = \frac{\sum_{p=1}^{P} \boldsymbol{B}_p^{f(c_k)}}{\sum_{p=1}^{P} \boldsymbol{M}_p^{s(c_k)}}. \tag{5.1}$$

Because our goal is to compare the Igh locus in pro-B and MEF, the normalizing factor obtained from the gene desert will be used to normalize all the ratios $r_{kp} = \frac{\boldsymbol{B}_p^{f(c_k)}}{\boldsymbol{M}_p^{s(c_k)}}$ for all pairs of primers $p = \{1, 2, \ldots, P_{Igh}\}$ in the Igh locus of replicate combination k. We will use log2-ratios to obtain a standard measure of how much the interactions in pro-B differ from those in MEF. The following equations show how to obtain a **corrected ratio** $\hat{r}_{kp}$

$$\hat{r}_{kp} = \frac{\boldsymbol{B}_p^{f(c_k)}}{\boldsymbol{M}_p^{s(c_k)}} \cdot \frac{1}{n_k}$$

$$log_2(\hat{r}_{kp}) = log_2(r_{kp}) - log_2(n_k)$$

where $log_2(\hat{r}_{kp})$ is the corrected log2-ratio of interactions between the cells for pair p and replicate combination k. As it will be shown in the next section, the computation of these factors needs to be done with primers that have the highest level of agreement between replicate combinations. Therefore, we implement an iterative process that filters out pairs of primers from the gene desert whose counts diverge between replicates. By keeping the most homogeneous primers we are able to compute more reliable normalizing factors.

5.7.6 Filtering primers in gene desert to improve the normalizing factors

We previously analyzed the gene desert for each combination of replicates separately and obtained a normalizing factor. Using our index notation, for $k = 1, c_k = (1,1)$ –replicate 1 of pro-B and replicate 1 of MEF– we obtained a normalizing factor n_1; for $k = 2, c_k = (2,2)$ we obtained n_2, and so forth.

Ideally, we would expect the corrected ratio $\hat{r}_{kp}$ to be very similar to the corrected ratio $\hat{r}_{qp}$, both for primer pair p, where $k, q \in K$ are different indices representing different combinations of replicates (Quackenbush, 2002; Geller et al., 2003). In order to improve the readability of the following equations, we assume $k = 1$ as the combination of replicates (1, 1) and $q = 2$ as (2, 2). The equations apply to any other combination of replicates.

In formal terms, for all primer pairs p and combinations of replicates 1 and 2:

$$\hat{r}_{1p} = \frac{\boldsymbol{B}_p^1}{\boldsymbol{M}_p^1} \cdot \frac{1}{n_1} \tag{5.2}$$

$$\hat{r}_{2p} = \frac{\boldsymbol{B}_p^2}{\boldsymbol{M}_p^2} \cdot \frac{1}{n_2} \tag{5.3}$$

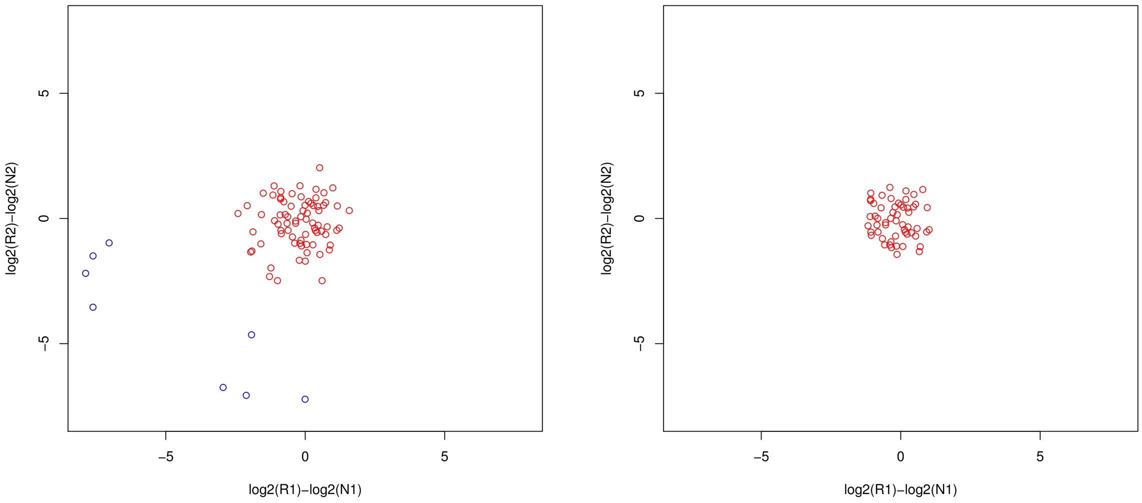
we expect

$$\frac{\hat{r}_{1p}}{\hat{r}_{2p}} = \frac{\boldsymbol{B}_p^1}{n_1 \boldsymbol{M}_p^1} \cdot \frac{n_2 \boldsymbol{M}_p^2}{\boldsymbol{B}_p^2} \sim 1$$

which is equivalent to

$$log_2(\frac{\hat{r}_{1p}}{\hat{r}_{2p}}) \sim 0$$

A scatter plot with these corrected ratios is shown in Figure 50(a). Some pairs of primers do not seem to agree between the two replicates (marked in blue as outliers with more than two standard deviations from the rest). These are the primers in the gene desert that we postulate should be excluded from the computation of the normalizing factors.



(a) Before running the iterative process.

(b) After six iterations all outliers are removed

Figure 50. Corrected log2-ratios in chr5. Combination of replicates (1,1) and (2,2). Outliers are marked in blue.

The algorithm we developed to filter pairs of primers used in the computation of normalizing factors, consists of the following steps:

1. For every possible combination of replicates k and q, compute log2-ratios of pro-B over MEF.
2. Exclude primer pairs that are more than 2 standard deviations away from the rest.
3. Compute the normalizing factors and correct counts of remaining primers.
4. Repeat from 1. until there are no more primers to discard.

After six iterations of our algorithm, the pairs of primers that show consistency in both combinations of replicates are marked in red and grouped around zero in Figure 50(b).

The normalizing factors we obtain after running the algorithm on all combinations of replicates are shown in Table XXXII.

TABLE XXXII

NORMALIZING FACTORS OBTAINED FROM ALL COMBINATIONS OF REPLICATES (WITH FILTERING).

Set	Cell/replicate	Cell/replicate	Coefficient
1	pro-B1	MEF1	0.779544
	pro-B2	MEF1	0.692332
2	pro-B1	MEF2	0.587048
	pro-B2	MEF1	0.709502
3	pro-B1	MEF1	0.779544
	pro-B2	MEF2	0.531143
4	pro-B1	MEF2	0.587162
	pro-B2	MEF2	0.531143

Finally, to compute the normalizing coefficient of a specific cell and replicate, we obtain an average of the factors in Table XXXII that include that replicate. In this way, we obtain:

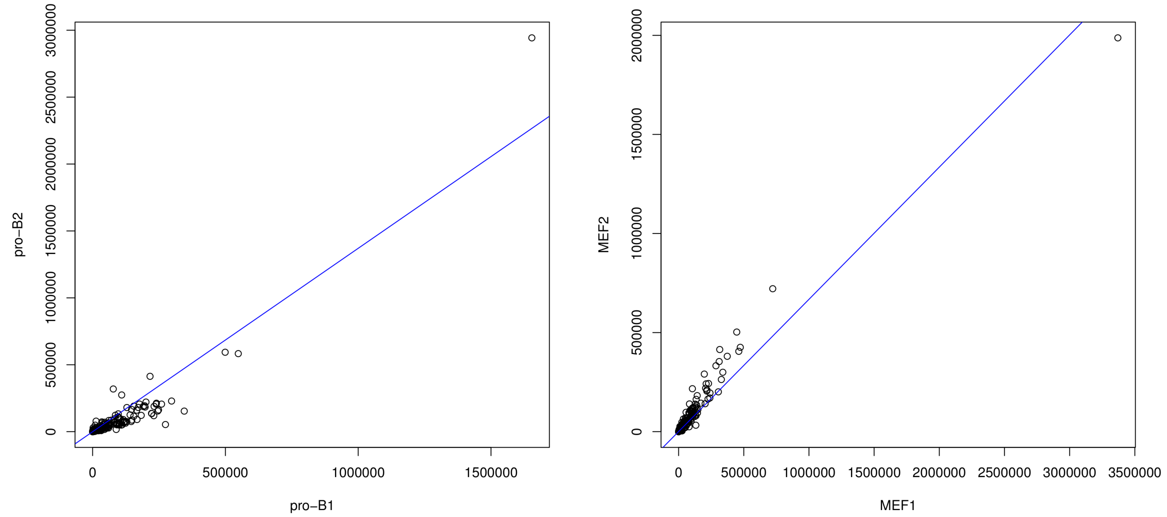
- Normalizing coefficient for MEF, rep. $1 = 0.7402305 = \frac{0.779544+0.692332+0.709502+0.779544}{4}$
- Normalizing coefficient for MEF, rep. $2 = 0.5591240 = \frac{0.587048+0.531143+0.587162+0.531143}{4}$

5.7.7 Normalizing counts and combining replicates in Igh locus

At this point, there is no further need to contemplate different combinations of replicates. After all, the normalizing coefficients we obtained in the previous step are for each specific replicate. Because our ultimate goal is to compare pro-B and MEF, in this section we describe the steps we followed to combine the replicates of the same cell type.

Before combining the replicates, we computed a correlation of the interaction counts between the replicates of each cell. Both, pro-B and MEF had high Spearman correlation coefficients: $\rho = 0.704$ for pro-B and $\rho = 0.885$ for MEF and statistically significant p-values $<$ 2.2e-16. Figure 51 shows the scatter plots of the interactions of replicate 1 (X-axis) versus replicate 2 (Y-axis).

Based on the fact that there is a strong agreement between replicates, we decided to combine them. To that effect, we started by computing an average of the corrected log2-ratios as shown below.



(a) In pro-B, replicate 1 vs. 2, $\rho = 0.704$ and p-value < 2.2e-16

(b) In MEF, replicate 1 vs. 2, $\rho = 0.885$ and p-value < 2.2e-16

Figure 51. Spearman correlation of interaction counts in both replicates of the same cell.

For every pair of primers p we have:

$$\begin{aligned} v_p &= \frac{log_2(\hat{r}_{1p}) + log_2(\hat{r}_{2p})}{2} \\ &= \frac{1}{2} log_2(\hat{r}_{1p} \cdot \hat{r}_{2p}) \\ &= log_2(\sqrt{\hat{r}_{1p} \cdot \hat{r}_{2p}}) \end{aligned}$$

substituting from Equations 5.2 and 5.3

$$\begin{aligned} v_p &= log_2(\sqrt{\frac{\boldsymbol{B}_p^1}{n_1 \boldsymbol{M}_p^1} \cdot \frac{\boldsymbol{B}_p^2}{n_2 \boldsymbol{M}_p^2}}) \\ &= log_2(\sqrt{\frac{\boldsymbol{B}_p^1 \boldsymbol{B}_p^2}{\boldsymbol{M}_p^1 \boldsymbol{M}_p^2} \cdot \frac{1}{n_1 n_2}}) \end{aligned}$$

and the normalized average value for pair p is

$$v_p = log_2(\sqrt{\frac{\boldsymbol{B}_p^1\boldsymbol{B}_p^2}{\boldsymbol{M}_p^1\boldsymbol{M}_p^2}}) - log_2(\sqrt{n_1 n_2}) \tag{5.4}$$

In Equation 5.4 we see that the normalized average value v_p for pair p is obtained by computing the geometric mean of interactions in pro-B ($\sqrt{\boldsymbol{B}_p^1\boldsymbol{B}_p^2}$) and MEF ($\sqrt{\boldsymbol{M}_p^1\boldsymbol{M}_p^2}$) and then correcting with the normalizing coefficients n_1 and n_2 computed for both replicates of MEF in Section 5.7.6. Equation 5.4 generalizes very easily to an experiment with m replicates where the normalized average value for the primer pair p is:

$$v_p = log_2\left(\sqrt[m]{\prod_{i=1}^{m}\frac{\boldsymbol{B}_p^i}{\boldsymbol{M}_p^i}}\right) - log_2\left(\sqrt[m]{\prod_{i=1}^{m} n_i}\right) \tag{5.5}$$

There are a few special considerations with regards to Equation 5.5 that need to be made in order to avoid singularities in the operations:

- If $\exists\ i \in \{1, 2, \ldots m\} \mid \boldsymbol{B}_p^i = 0$ then set $\boldsymbol{B}_p^i = 0.01$. The same applies to $\boldsymbol{M}_p^i$.
- $\forall\ i \in \{1, 2, \ldots m\} \mid \boldsymbol{B}_p^i = 0 \wedge \boldsymbol{M}_p^i = 0$ set $v_p = 0$.

At the end of this step, we create two average-normalized matrices: $\boldsymbol{\mathcal{M}}_{norm}^{proB}$ and $\boldsymbol{\mathcal{M}}_{norm}^{MEF}$ that will contain the normalized interaction counts of all pairs of primers, for the combined replicates in pro-B and MEF respectively.

5.8 Visualizing normalized data

After implementing the methods described in the previous sections we were able to scale the interactions in the Igh locus and to combine both replicates of each cell type. Here we obtain a first view of the binned contact maps for pro-B and MEF. In both cases, the data were binned using a bin size=100 Kb and a bin step=10 Kb. Figure 52 shows the binned contact map of the matrix $\boldsymbol{\mathcal{M}}_{norm}^{proB}$ computed before and Figure 53 does the same for the matrix $\boldsymbol{\mathcal{M}}_{norm}^{MEF}$.

At first sight, both cells seem to have the same pattern of interactions. Yet, there are noticeable differences that can be appreciated by looking more closely at the binned contact maps. For example, the genomic region between the coordinates 114,455,800 and 114,459,771 (between the 2nd and 3rd ticks in the Y-axis, from the topmost left corner) corresponds to an enhancer known as 3' regulatory region (3'RR). This region has been well-characterized in the mouse and it has been found to have a prominent role in the production of immunoglobulins in B cells (Rouaud et al., 2013).

From a biological standpoint, we expect to see more interactions around that region in pro-B than MEF, and the binned contact maps seems to validate it. Nevertheless, it is clear from these figures that we need a different comparison approach to determine what interactions are more prevalent in one cell than in the other.

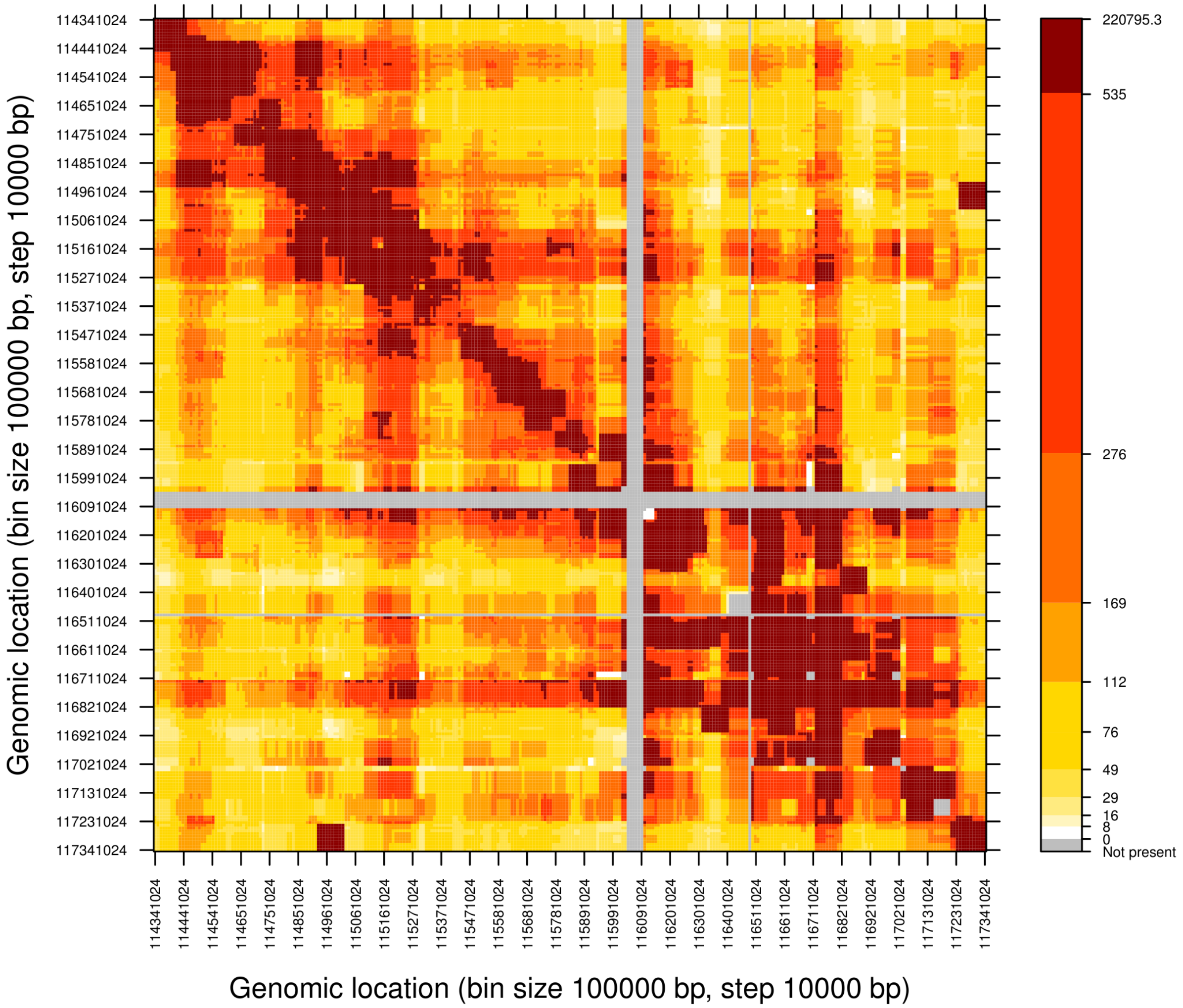


Figure 52. pro-B: binned contact map of Igh locus. Bin size=100 Kb, bin step=10 Kb

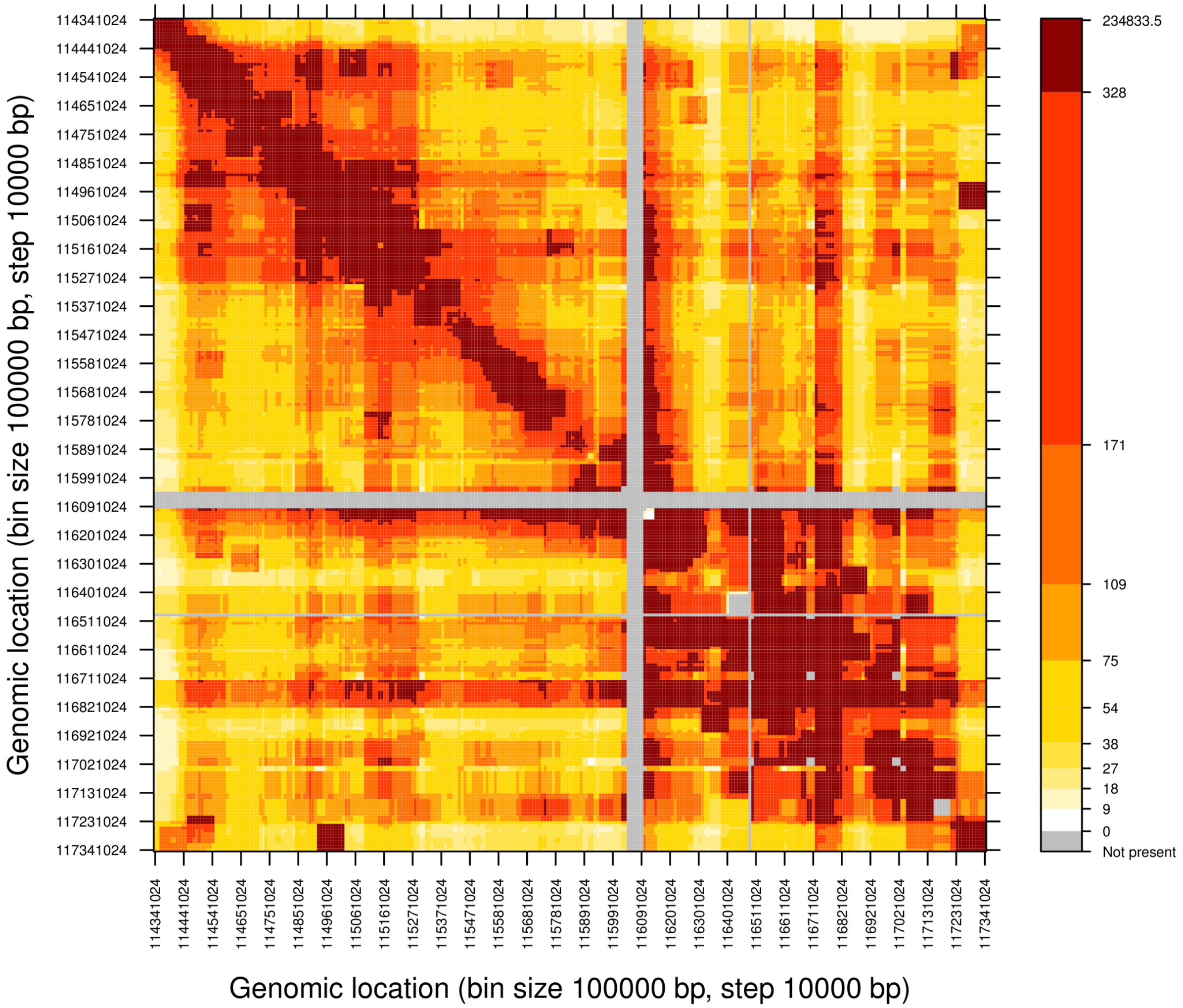


Figure 53. MEF: binned contact map of Igh locus. Bin size=100 Kb, bin step=10 Kb

5.9 Expected number of interactions

One final step before we can compare pro-B and MEF consists in removing the **expected interactions** between any pair of primers i and j. The input to this step are the two matrices $\boldsymbol{\mathcal{M}}_{norm}^{proB}$ and $\boldsymbol{\mathcal{M}}_{norm}^{MEF}$ computed at the end of Section 5.7.7. The computation of expected interactions can be better explained as a noise reduction step that aims at keeping only the true interactions between i and j, eliminating those that arise by chance. Of course, this is easier said than done. The question we face is: What is a *neutral reference state* that can be used to determine the expected number of interactions between any primers i and j?

In reality, there is no answer to the previous question and many researchers have attempted to approximate the concept of *expected* through different methods. The most popular approach creates a set of points $\mathcal{P} = \{(x_{11}, y_{11}), (x_{12}, y_{12}), \ldots, (x_{ij}, y_{ij}), \ldots, (x_{FR}, y_{FR})\}$ where F if the number of forward primers and R is the number of reverse primers; x_{ij} is the genomic distance between primers i and j; and y_{ij} is the number of interactions between primers i and j. These points are smoothed with a LOESS curve which in turn is used to predict the expected number of interactions between i and j (Lajoie et al., 2009; Baù et al., 2011). A different approach bases the calculation of expected contacts on the analysis of polymer models (Lieberman-Aiden et al., 2009). In these models, it has been observed that the number of interactions between two genomic regions decreases as the distance between the regions increases, thus following a power law $y = x^{-c}$ where the coefficient c has been estimated to be approximately -1.08. A recently released method to determine expected interactions models them using a Weibull distribution and implements the stochastic gradient descent method to learn the parameters of the distribution from the data (Phillips-Cremins et al., 2013).

All of these methods make assumptions about the data and/or about the underlying structure of chromatin and, unfortunately, there is no objective set of measures that can be used to determine which one is better.

For our analysis, we decided to estimate the expected interactions using LOESS smoothing (Cleveland, 1979; Cleveland and Devlin, 1988). In its first iteration of the paper, the author lays the foundations of the method (originally referred to as LOWESS for **LO**cally **WE**ighted **S**catterplot **S**moothing). The steps are (Cleveland, 1979):

1. For each x_{ij} define weights $w(x_{ij})$ using a weight function W which essentially gives more weight to data points nearer to the point where the estimation will be done.

2. For a point q in $\mathcal{P}$, compute a set of estimates $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, ..., \gamma_d)$ which will be parameters of a polynomial regression of degree d that minimize:

$$\sum_{k=1}^{|\mathcal{P}|} W(x_q, x_k)(y_k - \gamma_0 - \sum_{r=1}^{d} x_k \gamma_r)^2 \tag{5.6}$$

   where q is our estimating point; and $W(x_q, x_k)$ computes the weight between q and any other point k in $\mathcal{P}$.

3. Finally, the smoothed value is computed as:

$$\hat{y}_q = \gamma_0 + \sum_{r=1}^{d} x_q^r \gamma_r \tag{5.7}$$

where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, ..., \gamma_d)$ are the optimally estimated parameters in Equation 5.6.

In the second iteration of their work (Cleveland and Devlin, 1988), the method was renamed to LOESS for **LO**cally weighted regr**ESS**ion and the weight function W acquired a new parameter α. This parameter is used to determine the fraction of points –or span– in the vicinity of the estimating point that will be used to compute the weight. The larger the α, the stronger the smoothing effect of the overall function. A smaller α, on the other hand, forces the function to use few neighboring points exacerbating the locality of the estimation and thus creating a non-smooth curve that jitters between points.

We used LOESS with a second-degree polynomial ($d = 2$) and a fraction of points $\alpha = 0.1$. The LOESS curve computed for pro-B can be seen in Figure 54. Refer to the Appendix, where Figure 73 shows a LOESS curve for MEF. For this analysis, we excluded primers whose distance was less than 12 Kb as these are considered proximity events.

Figure 54 illustrates the behavior to which we alluded before when we briefly mentioned polymer models: the larger the distance between primers, the smaller the number of interactions we expect to find.

The LOESS curve was ultimately used to compute a new set of points $\hat{\mathcal{P}} = \{(x_{11}, \hat{y}_{11}), (x_{12}, \hat{y}_{12}), \ldots, (x_{ij}, \hat{y}_{ij}), \ldots, (x_{FR}, \hat{y}_{FR})\}$ where each $\hat{y}_{ij}$ represents the expected number of interactions between primers i and j. We obtained two sets of points, one for pro-B and one for MEF. With these sets of points we constructed two matrices containing the expected interactions for every pair of primers as computed by LOESS: $\boldsymbol{\mathcal{M}}^{proB}_{expected}$ for pro-B and $\boldsymbol{\mathcal{M}}^{MEF}_{expected}$ for

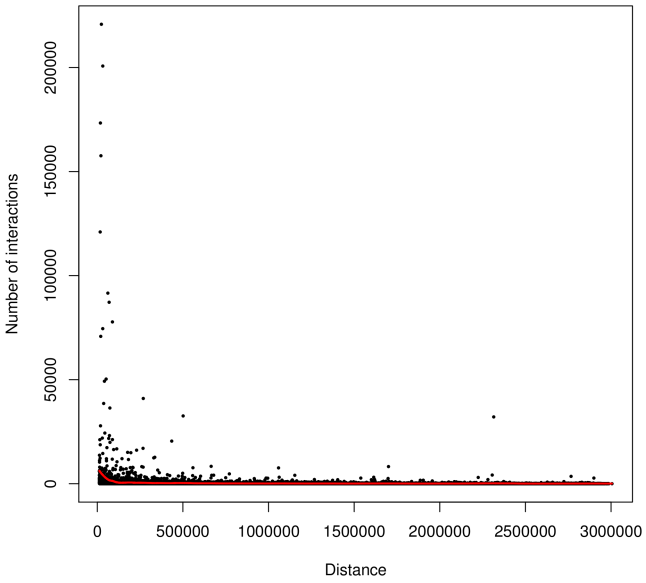


Figure 54. Interactions by distance in pro-B with LOESS curve ($d = 2$ and $\alpha = 0.1$)

MEF. The binned expected interactions of $\boldsymbol{\mathcal{M}}^{proB}_{expected}$ can be seen in Figure 55. Refer to the Appendix, Figure 74 for a similar binned contact map of MEF.

Upon closer inspection of Figure 55 reveals some interesting aspects about the expected values in pro-B. Firstly, we see that the highest number of expected interactions cluster around the diagonal. The diagonal –or its immediate neighborhood– represents the interactions between proximal fragments. As was mentioned before, contiguous regions of chromatin are more likely to interact with each other than with regions located far away. The farther we move from the diagonal, the more prominent the decrease in expected interactions is. For example, two fragments located around 114,441,024 (2nd tick in the X- and Y-axis, from topmost left) have

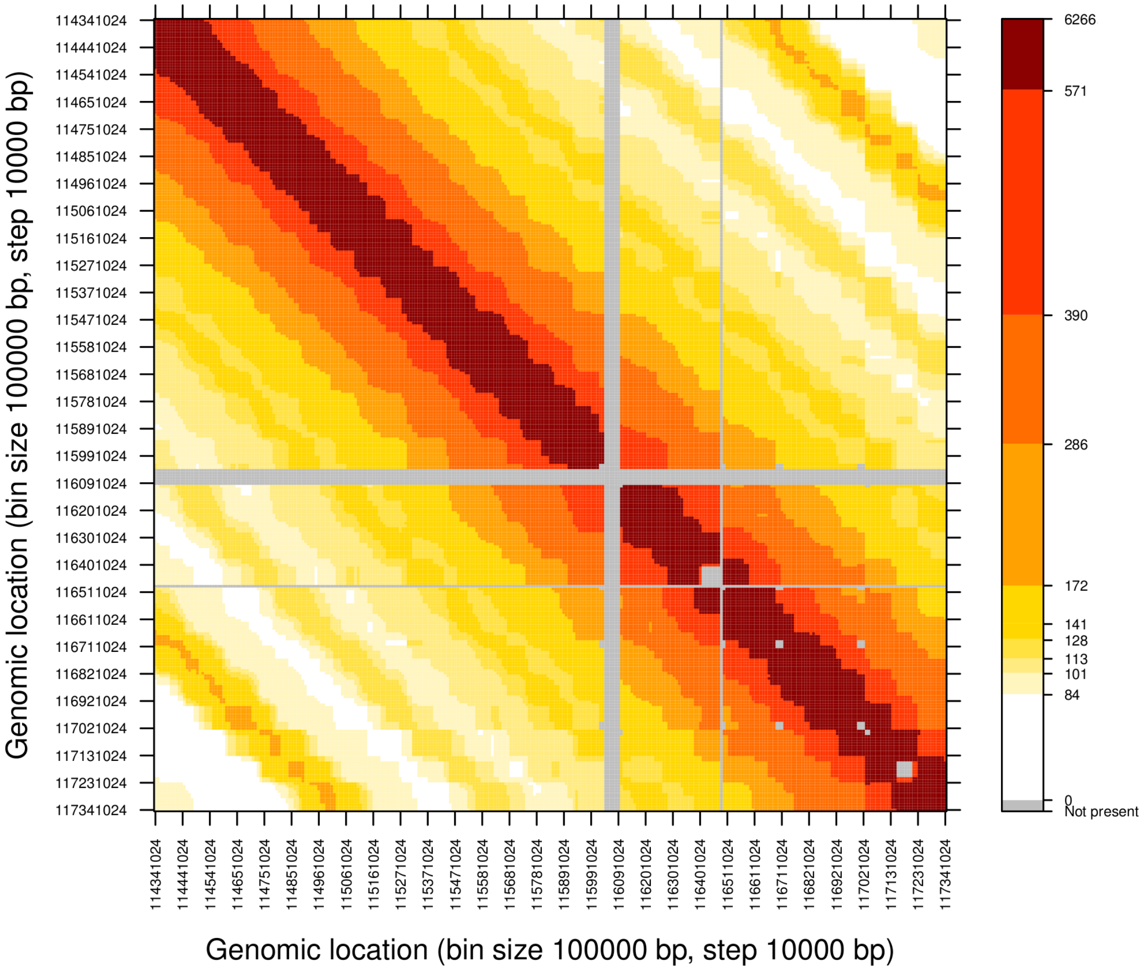


Figure 55. pro-B: binned expected interactions in Igh locus using LOESS. Bin size=100 Kb, bin step=10 Kb

an expected number of interactions that ranges between 571 and 6,266. Now, the expected interactions between one of those fragments and another fragment at 115,781,024 (15th tick from the left in the X-axis) is in the range of 84 to 101.

It is worth noting that, following the last example, as we move further to the right in the X-axis we see that the number of expected interactions increases again. This is an artefact of our 5C data and of the smoothing conducted by LOESS. It is simply saying that in our locus of interest, due to the layout of primers and the interactions they capture, there is a higher likelihood of finding long-range interactions of certain length (approximately, of 2.48 Mb long) vs. other long-range interactions (approximately, of 1.86 Mb long).

5.10 Comparing interactions in pro-B versus MEF

Finally, we are ready to compare prob-B and MEF. We first subtracted $\boldsymbol{\mathcal{M}}_{expected}^{proB}$ from $\boldsymbol{\mathcal{M}}_{norm}^{proB}$ and set to zero the negative counts (these represent the cases where the number of observed interactions is less than the expected). We repeated this process for MEF and created a binned contact map of pro-B over MEF with the log2-ratios of interactions. The binned contact map in Figure 56 shows, in shades of blue, regions where there are more interactions in MEF than pro-B. Conversely, shades of red show interactions that are more prevalent in pro-B than MEF. White indicates either absence of primers or a log2-ratio close to zero.

We can decouple Figure 56 by looking separately at the interactions that are either prevalent in pro-B or MEF. In order to achieve this we can set a threshold and retain only the interactions in one cell that, when compared to the other cell, are above certain threshold.

In our analysis of the Igh locus, we set a threshold of 1.6 log2-fold change (equivalent to a 3-fold difference in interaction counts) and reported the primers in pro-B that observed a log2-fold change difference with respect to MEF greater than the threshold. The interactions captured by these primers are then considered to be **prevalent** in pro-B. Figure 57 shows a

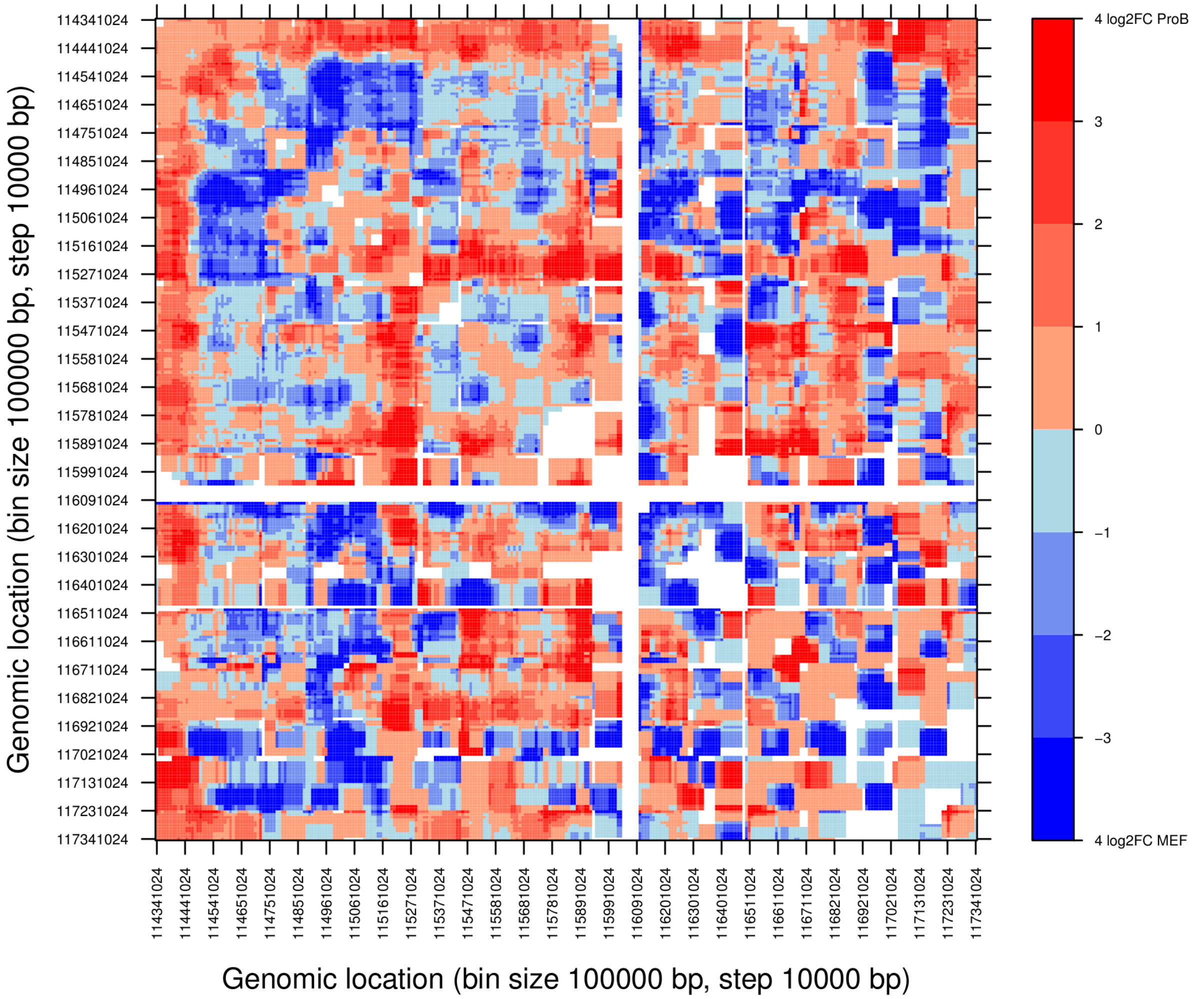


Figure 56. Binned contact map of log2-ratios: pro-B vs. MEF.

binned contact map of the interactions prevalent in pro-B. Similarly, we repeated this analysis looking for interactions prevalent in MEF and Figure 58 shows these interactions on the same scale as those in pro-B.

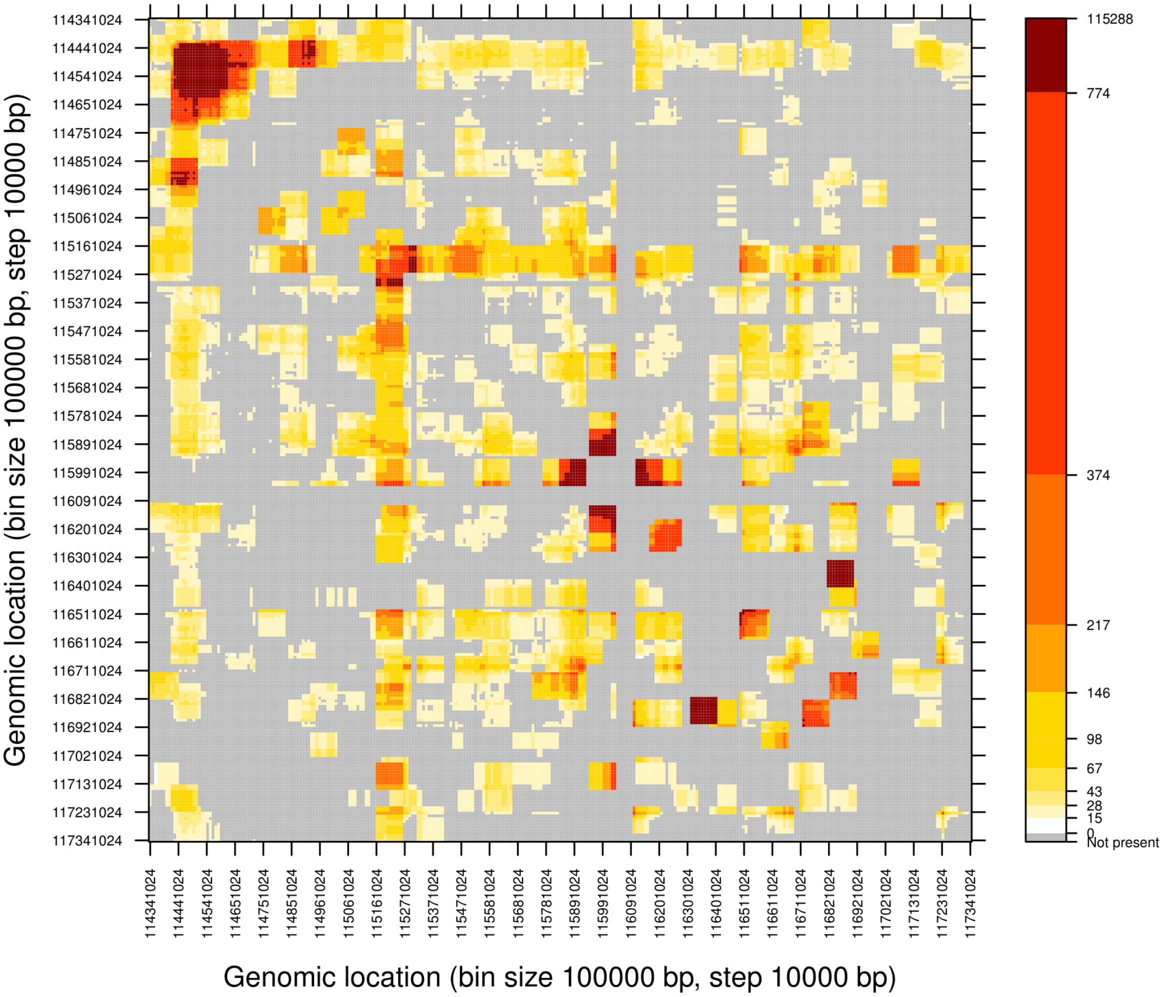


Figure 57. Binned contact map of interactions that are prevalent in pro-B, i.e., interactions of primers with more than a 3-fold change difference in pro-B with respect to MEF. Bin size=100 Kb, bin step=10 Kb

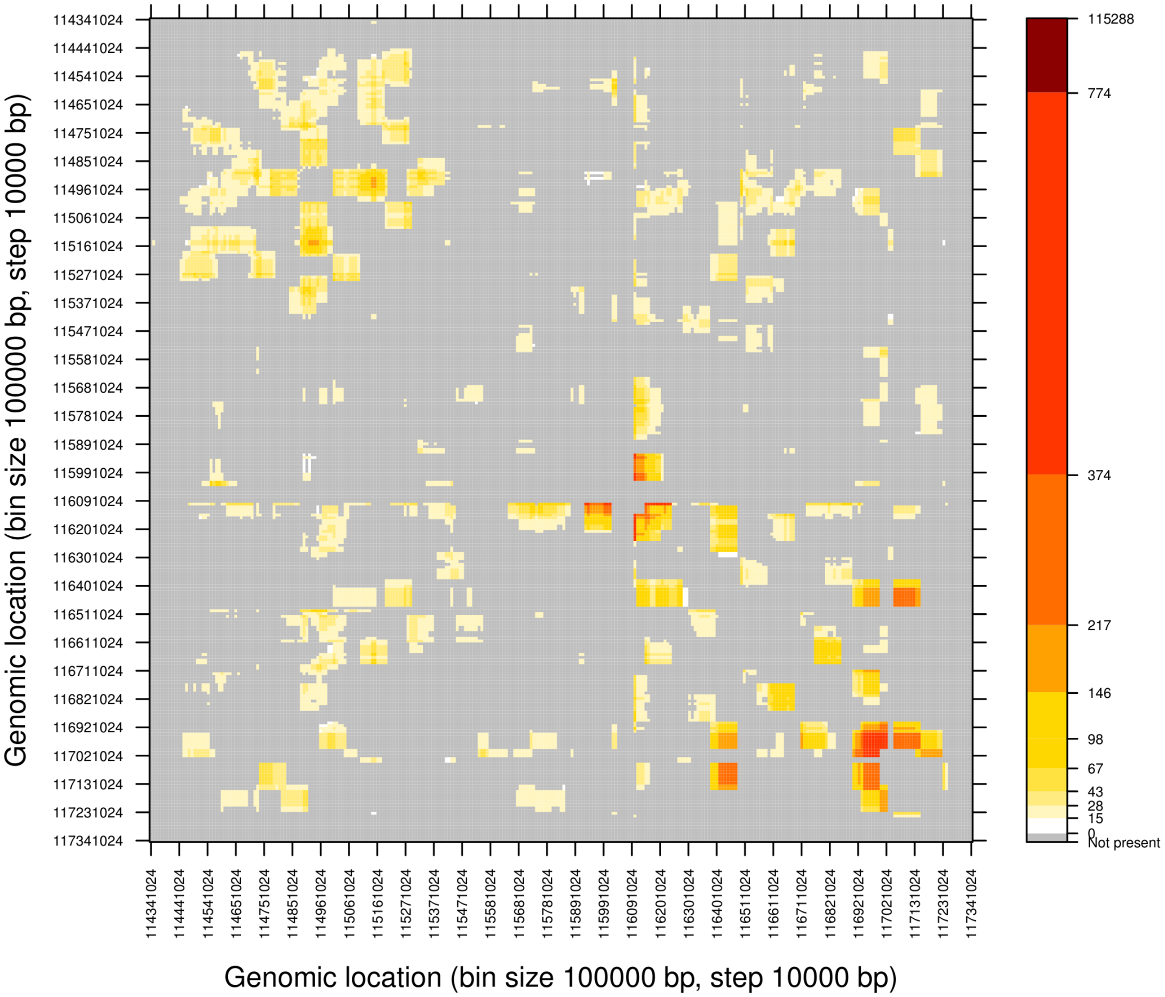


Figure 58. Binned contact map of interactions that are prevalent in MEF, i.e., interactions of primers with more than a 3-fold change difference in MEF with respect to pro-B. Bin size=100 Kb, bin step=10 Kb

The interactions reported in Figure 57 are of particular importance since they show the genomic regions of the Igh locus that undergo gene rearrangement prior to the production of immunoglobulins. These long-range interactions are pro-B specific and provide a clear picture of the conformational structure of the locus in this type of cell.

5.11 Pipeline implementation

All the analyses conducted in this chapter were performed with an in-house pipeline developed in R and Python. The mapping of NGS data was done with `Bowtie` (Langmead et al., 2009). The contact maps and other visualization tools were created using the R packages `lattice` and `gplots`.

The source code of the pipeline was donated to the Research Resources Center at the University of Illinois-Chicago with the sole purpose that future 5C projects conducted at the university would benefit from the methods and algorithms we developed.

5.12 Conclusion

In this chapter we have focused on the development of methods and tools to manipulate chromosome conformation data. Firstly, we provided technical details of the raw experimental data that are generated by a 5C experiment. We illustrated with step-by-step examples the characteristics of the visualization methods we developed. We showed in clear terms the methods to transform raw experimental data –representing chromatin interactions– into coherent and noise-free visual representations of those interactions.

With the ultimate goal of comparing interaction data from two cell types: pro-B and MEF, we proposed a normalization method to scale and combine the interactions in different replicates. Additionally, we implemented a noise reduction method based on LOESS smoothing to eliminate interactions that would normally arise due to the proximity of primers.

Lastly, we combined all the pieces of the puzzle to obtain a binned contact map describing the regions of the Igh locus where interactions are more prevalent in pro-B versus MEF, and vice versa.

CHAPTER 6

DISCUSSION

In hopes of gaining a better understanding of how genes are regulated, the work we presented has made use of different high-throughput technologies and has required the implementation of different mathematical models and statistical methods. This integrative and multidisciplinary approach has been the guiding compass of our research work.

Our analysis of chromosome conformation is an example of true multidisciplinary work. Starting with the burdensome –yet essential– pre-processing of data, and all the way to the creation of visualization tools for the results, it was the close collaboration with domain experts that helped shape our proposed methods.

In regards to data integration, we illustrated an example of it with our probabilistic framework to infer regulators of pathways. Here we combined seemingly disparate data sources such as binding predictions of TFs and microRNAs, the structure of molecular pathways, and mRNA/microRNA expression profiles. Each of these data sources attempted to predict the mechanisms of gene regulation from different perspectives, but it was solely through data integration that our predictions became more grounded and less biased.

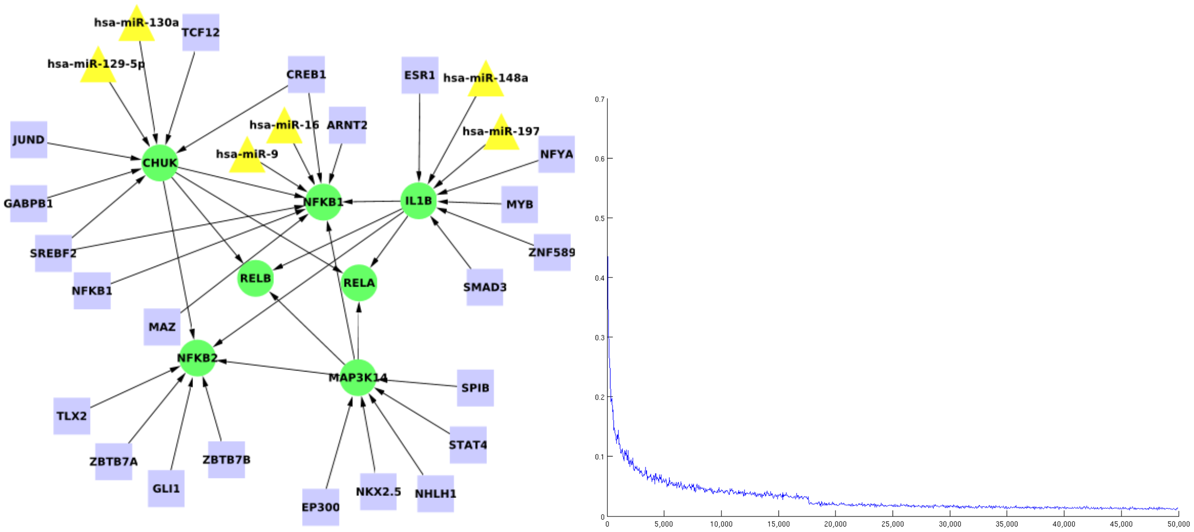
Another example that reflects our true integrative philosophy is the analysis pipeline we developed for methylation and mRNA datasets. With the help of sound mathematical models we were able to equate changes of gene expression with the amount of methylation located at different regions of a gene.

It is our personal belief that more methods based on data integration must be developed in order to extract biologically relevant information from the ever increasing abundance of biological data. Therein lies the greatest bioinformatics and machine learning challenge for the future.

APPENDICES

Appendix A

SUPPLEMENTARY INFORMATION FROM CHAPTER 3



(a) (b)

Figure 59. (a) Toy BN of 36 nodes and (b) its error in approximating marginals using Gibbs sampler.

TABLE XXXIII. STATISTICS OF NODES IN PATHWAYS

		Original pathway		Pre-processed pathway		Merged pathway		Transcription Factors targeting nodes in the pathway					microRNAs targeting nodes in the pathway				
KEGG Id	Name	Nodes	Edges	Nodes	Edges	Nodes	Edges	Nodes with TF	Avg	Min	Max	Median	Nodes with miR	Avg	Min	Max	Median
4010	MAPK signaling pathway	268	891	227	482	524	2019	221	95.3	32	165	136	183	10.7	1	54	9
4012	ErbB signaling pathway	87	214	77	171	287	703	76	92.6	34	173	58	63	14.0	1	68	1
4020	Calcium signaling pathway	177	505	154	456	406	1503	154	90.4	34	161	44	118	10.2	1	42	19
4060	Cytokine-cytokine interaction	265	254	217	187	470	1397	199	88.6	32	164	91	102	9.6	1	68	7
4070	Phosphatidylinositol signaling	78	1402	69	10	267	484	68	94.8	34	163	132	58	10.3	1	40	9
4080	Neuroactive interactions	272	46	237	34	499	1483	227	86.5	33	154	67	144	7.8	1	37	2
4110	Cell cycle	128	680	109	1011	335	1718	103	98.2	32	161	76	72	12.5	1	50	1
4114	Oocyte meiosis	114	458	88	682	297	1288	86	91.6	32	154	97	62	14.5	1	59	12
4115	p53 signaling pathway	69	86	58	68	225	424	52	91.8	44	151	132	39	14.2	1	59	2
4122	Sulfur relay system	10	9	7	7	47	53	7	97.6	53	121	121	5	2.8	1	6	6
4130	SNARE interactions	36	46	30	39	157	237	29	92.0	40	144	72	24	10.9	1	37	2
4140	Regulation of autophagy	34	7	24	6	119	144	24	84.2	36	124	74	12	9.8	1	29	3
4141	Protein processing in ER	167	340	137	292	387	1201	134	91.5	27	156	116	107	8.4	1	42	31
4150	mTOR signaling pathway	52	80	44	55	200	378	43	90.3	32	152	94	37	17.1	2	59	23
4210	Apoptosis	86	169	76	159	256	598	72	93.9	36	171	67	42	12.1	1	36	3
4310	Wnt signaling pathway	151	775	129	863	391	1778	125	95.1	30	161	86	113	12.5	1	62	14
4330	Notch signaling pathway	47	138	34	94	171	316	32	92.9	27	131	72	27	10.3	1	26	15
4340	Hedgehog signaling pathway	56	145	43	108	170	411	42	93.8	44	160	132	37	10.2	1	47	6
4350	TGF-beta signaling pathway	85	226	76	188	275	714	73	90.5	42	161	99	61	14.9	1	68	13
5200	Pathways in cancer	327	1104	295	1120	595	3121	283	95.8	32	164	59	243	11.3	1	59	3
5210	Colorectal cancer	62	104	59	77	240	464	57	95.3	42	161	81	46	12.8	1	47	8
5211	Renal cell carcinoma	70	109	64	103	249	558	60	95.6	32	164	77	56	12.4	1	39	9
5212	Pancreatic cancer	70	137	67	133	248	580	65	96.5	32	161	63	51	14.8	1	50	2
5213	Endometrial cancer	52	87	50	69	218	417	49	93.1	42	165	78	40	12.7	1	42	28
5214	Glioma	65	189	59	186	247	599	57	95.1	34	164	74	48	18.9	1	59	13
5215	Prostate cancer	89	295	81	308	292	863	80	95.2	41	164	76	68	14.9	1	59	4
5216	Thyroid cancer	29	49	29	49	153	265	29	92.1	42	163	63	26	10.9	1	42	2
5217	Basal cell carcinoma	55	310	45	382	211	708	43	92.9	32	160	77	42	8.4	1	37	9
5218	Melanoma	71	281	66	227	262	682	64	94.0	42	164	72	50	16.7	1	59	8
5219	Bladder cancer	42	46	41	46	181	333	40	92.8	32	151	71	32	12.7	1	47	10
5220	Chronic myeloid leukemia	73	185	69	155	260	612	67	103.8	42	161	110	54	16.8	1	50	21
5221	Acute myeloid leukemia	58	155	54	131	235	517	53	96.9	42	156	137	45	12.8	1	42	12
5222	Small cell lung cancer	85	225	80	186	279	715	77	93.1	35	152	67	60	13.0	1	50	6
5223	Non-small cell lung cancer	54	124	52	88	227	443	50	97.9	34	152	123	41	16.4	1	50	12

TABLE XXXIV

SELECTED MARGINALS FOR THE CELL CYCLE PATHWAY (6 DATASETS + ENERLY)

Node	Marginals									
	Scenario #1					Scenario #2				
	very low	medium low	medium	medium high	very high	very low	medium low	medium	medium high	very high
SMAD3	**0.14**	**0.81**	0.05	0	0	0.2	0.63	0.08	0.03	0.06
CDKN2B	0.96	0.01	0.01	0.01	0.01	1	0	0	0	0
CCND1			**(de)**			**(de)**				
CDK4	0.01	0.01	0	0.97	0.02	0	0	0	0.68	0.32
TFE3			0.64	0.36				0.86	0.13	0.01
LMO2		0.02	0.92	0.06			0.02	0.93	0.06	
ELK4	0.99	0.01	0.01			0.97	0.01		0.01	0.01
SREBF2			1.0			0.02		0.98		
PAX4	1.0					1.0				
NFIC	0.12	0.34	0.47	0.07		0.2	0.33	0.41	0.06	
STAT6			0.07	0.94				0.08	0.91	0.01
SREBF1			0.01	0.99	0.01			0.01	0.99	0.01
NFIB	0.1	0.27	0.28	0.26	0.09	0.12	0.23	0.33	0.23	0.1
PPARA	0.98	0.01		0.01		1.0				
hsa-mir-375	0.3	0.18	0.16	0.12	0.24	0.37	0.19	0.14	0.13	0.18

Appendix A (Continued)

TABLE XXXV

SELECTED MARGINALS FOR THE P53 SIGNALING PATHWAY (6 DATASETS + BUFFA)

Node	Marginals									
	Scenario #1					Scenario #2				
	very low	medium low	medium	medium high	very high	very low	medium low	medium	medium high	very high
STAT5B	0.24	0.74	0.02	0.01		0.12	0.88	0.01		
PERP	**(de)**							**(de)**		
IGFBP3	0.2	0.18	0.21	0.17	0.23	0.15	0.2	0.23	0.23	0.19

TABLE XXXVI

SELECTED MARGINALS FOR THE MAPK SIGNALING PATHWAY (6 DATASETS + BUFFA)

Node	Marginals									
	Scenario #1					Scenario #2				
	very low	medium low	medium	medium high	very high	very low	medium low	medium	medium high	very high
STAT5B	0.35	0.53	**0.08**	**0.04**	**0.01**	0.23	0.68	0.05	0.04	
MAP3K12					**(de)**		**(de)**			
NFKB2	0.18	0.21	0.2	0.23	0.18	0.24	0.16	0.18	0.19	0.22
RRAS2	0.17	**0.2**	**0.2**	**0.23**	**0.2**	0.83	0.03	0.1	0.03	0.02
PTPRR	0.18	0.15	0.24	0.23	0.19	0.17	0.19	0.23	0.2	0.2
FGF23	0.22	0.21	0.19	0.2	0.18	0.2	0.23	0.17	0.21	0.2
MAPK10	0.14	0.2	0.19	0.23	0.23	0.18	0.17	0.2	0.19	0.26

Appendix B

SUPPLEMENTARY INFORMATION FROM CHAPTER 4

B.1 Results for the Illumina 27K platform: a case study

In order to illustrate a different set of results that can be obtained using `me-mRNA-pipe`, we downloaded and processed another publicly available dataset that focused on the analysis of methylation patterns in different tissues and different species (Pai et al., 2011). In Pai's study, the authors analyzed three types of tissue: heart (H), liver (L) and kidney (K) in two species: humans (Hs) and chimpanzees (Pt). Their focus was to characterize how the observable differences in DNA methylation between species affected the interspecies differences in gene expression.

The methylation data were extracted from the same samples of a previous interspecies mRNA study (Blekhman et al., 2008). For the analysis of methylation profiles, the Illumina 27K array described in section 4.3.3 was used (GEO Series GSE37020). The analysis of the mRNA expression data was performed using a custom multi-primate Nimblegen 388K microarray (GEO Series GSE11560).

In contrast to the ambitious scope of the work in (Pai et al., 2011) and (Blekhman et al., 2008), we limited our analysis to the human samples and focused on identifying relationships between methylation and gene expression in the three different tissues. Due to the fact that `me-mRNA-pipe` compares only two phenotypes at a time, it was executed three times to compare:

i) kidney vs. heart, ii) kidney vs. liver, iii) heart vs. liver and three different output directories were created for the three runs. It is important to note that in Pai's study each sample had two technical replicates. In the same way that the authors did not merge the replicates, in our analysis we measured the contribution of each replicate independently of the other.

B.1.1 Data pre-processing

Our first step was to determine the quality of the data in all samples. We analyzed the hierarchical clustering and PCA plots obtained from each run and identified two methylation samples in kidney (HsK1a and HsK1b) that did not cluster together with samples of the same tissue. This can be seen in Figure 60 where the two problematic samples cluster together at the left of the figure, far from the other kidney samples. The PCA plots, before and after excluding the samples, are shown in Figure 61(a) and Figure 61(b) respectively. It should be noted that after excluding the two samples, the first principal component is able to explain 11% more of the variance compared to when all samples were included.

Appendix B (Continued)

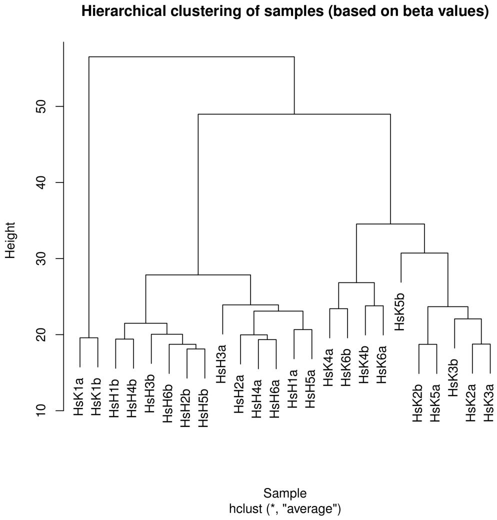
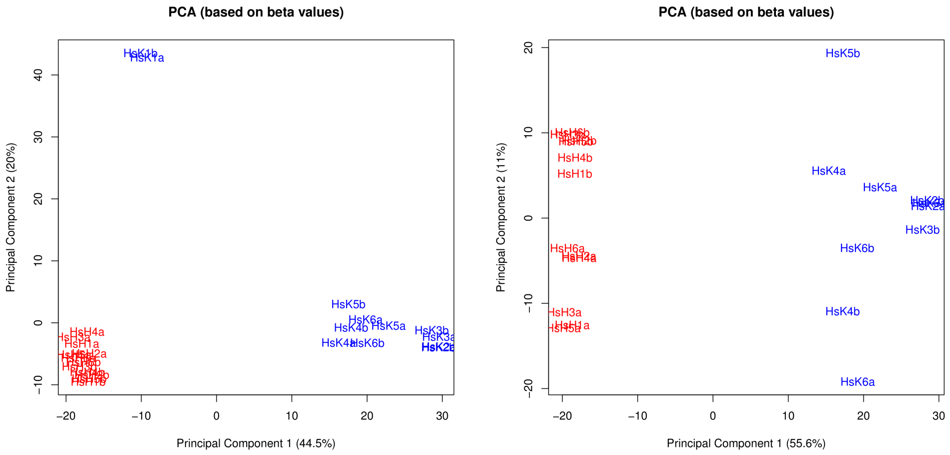


Figure 60. Hierarchical clustering of all samples in heart vs. kidney.

Appendix B (Continued)



(a) PCA of all samples

(b) After excluding HsK1a and HsK1b

Figure 61. Pre-processing of methylation data using PCA to identify samples that should be excluded from the analysis.

All methylation samples in Pai's study and their matched mRNA samples are listed in Table XXXVII. Only the samples of human tissues are shown in the table. Additionally, the two samples above mentioned and which are excluded from further analysis are greyed out.

Appendix B (Continued)

TABLE XXXVII

PHENOTYPE INFORMATION FOR METHYLATION SAMPLES IN PAI'S STUDY AND MRNA SAMPLES IN BLEKHMAN'S STUDY.

Methylation data			**mRNA data**		
Illumina 27K array			Nimblegen custom multi-primate 388K array		
Sample Id	GEO Id	Phenotype	Sample Id	GEO Id	Phenotype
HsL1a	GSM639308	LIVER	HsL1a	GSM291143	LIVER
HsL1b	GSM639309	LIVER	HsL1b	GSM291260	LIVER
HsL2a	GSM639310	LIVER	HsL2a	GSM291261	LIVER
HsL2b	GSM639311	LIVER	HsL2b	GSM291262	LIVER
HsL3a	GSM639312	LIVER	HsL3a	GSM291263	LIVER
HsL3b	GSM639313	LIVER	HsL3b	GSM291266	LIVER
HsL4a	GSM639314	LIVER	HsL4a	GSM291269	LIVER
HsL4b	GSM639315	LIVER	HsL4b	GSM291270	LIVER
HsL5a	GSM639316	LIVER	HsL5a	GSM291271	LIVER
HsL5b	GSM639317	LIVER	HsL5b	GSM291272	LIVER
HsL6a	GSM639318	LIVER	HsL6a	GSM291274	LIVER
HsL6b	GSM639319	LIVER	HsL6b	GSM291279	LIVER
HsK1a	GSM639332	KIDNEY	HsK1a	GSM291972	KIDNEY
HsK1b	GSM639333	KIDNEY	HsK1b	GSM291973	KIDNEY
HsK2a	GSM639334	KIDNEY	HsK2a	GSM291975	KIDNEY
HsK2b	GSM639335	KIDNEY	HsK2b	GSM291976	KIDNEY
HsK3a	GSM639336	KIDNEY	HsK3a	GSM291977	KIDNEY
HsK3b	GSM639337	KIDNEY	HsK3b	GSM291978	KIDNEY
HsK4a	GSM639338	KIDNEY	HsK4a	GSM291979	KIDNEY
HsK4b	GSM639339	KIDNEY	HsK4b	GSM291980	KIDNEY
HsK5a	GSM639340	KIDNEY	HsK5a	GSM291981	KIDNEY
HsK5b	GSM639341	KIDNEY	HsK5b	GSM291982	KIDNEY
HsK6a	GSM639342	KIDNEY	HsK6a	GSM291983	KIDNEY
HsK6b	GSM639343	KIDNEY	HsK6b	GSM291984	KIDNEY
HsH1a	GSM639355	HEART	HsH1a	GSM292009	HEART
HsH1b	GSM639356	HEART	HsH1b	GSM292010	HEART
HsH2a	GSM639357	HEART	HsH2a	GSM292011	HEART
HsH2b	GSM639358	HEART	HsH2b	GSM292012	HEART
HsH3a	GSM639359	HEART	HsH3a	GSM292013	HEART
HsH3b	GSM639360	HEART	HsH3b	GSM292014	HEART
HsH4a	GSM639361	HEART	HsH4a	GSM292015	HEART
HsH4b	GSM639362	HEART	HsH4b	GSM292016	HEART
HsH5a	GSM639363	HEART	HsH5a	GSM292017	HEART
HsH5b	GSM639364	HEART	HsH5b	GSM292018	HEART
HsH6a	GSM639365	HEART	HsH6a	GSM292019	HEART
HsH6b	GSM639366	HEART	HsH6b	GSM292020	HEART

In our analysis, mRNA probes were considered to be differentially expressed if the adjusted *p*-value of their difference between phenotypes was less than 0.01 (See section 4.4.3 for more details). On the other hand, DNA methylation probes were considered to be differentially methylated if the difference in beta values (Δbeta) was greater than 0.1. Table XXXVIII shows a summary of the number of differentially expressed mRNAs, the number of differentially methylated CpGs and their overlap computed by `me-mRNA-pipe`.

TABLE XXXVIII

SUMMARY OF DIFFERENTIALLY METHYLATED CPGS AND THEIR OVERLAP WITH DIFFERENTIALLY EXPRESSED GENES.

	Tissue comparison					
	Kidney vs. Heart		**Kidney vs. Liver**		**Heart vs. Liver**	
	count	**%**	**count**	**%**	**count**	**%**
Number of differentially methylated CpGs (DMCpG)	3,161		3,708		4,503	
Number of differentially expressed RefSeqIds (DER)	11,186		10,662		11,851	
▷ DMCpGs that overlap with at least one DER	1,692	53.5%	1,986	53.6%	2,657	59.0%
▷ DERs that overlap with at least one DMCpG	1,398	12.5%	1,597	15.0%	2,115	17.8%

B.1.2 ANOVA analysis

Due to the fact that the Illumina 27K array has few CpGs in the promoter of genes, the overlap between differentially methylated CpGs and differentially expressed genes does not leave too much room for an extensive ANOVA analysis. If we refer to the heart vs. liver comparison, Figure 62 shows a large number of differentially expressed genes (> 1,500) having only one differentially methylated CpG associated to them. From Table XXXVIII we have 2,115 differentially expressed genes that overlap with at least one differentially methylated CpG. Of these genes, 1,892 (89.5%) have only one overlapping CpG or, when they have more than one, they are in the same location. Even so, our ANOVA analysis reported some interesting results.

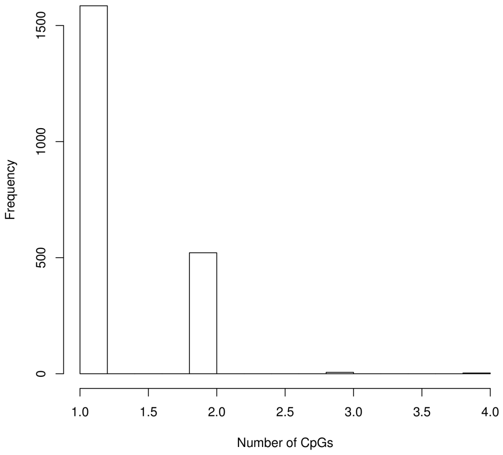


Figure 62. Histogram of the overlap of differentially methylated CpGs with differentially expressed genes.

As we did in Section 4.5.2, the top 20 genes from each tissue comparison were submitted to DAVID in search of GO term enrichment. Table XXXIX shows the enriched elements with an FDR $< 10\%$.

The GO terms enriched in the kidney vs. liver analysis are particularly interesting since they refer to inflammation and wound-healing response. In liver tissue, fibrosis is a response to wounding that involves different cells and mediators to contain the injury and it has been associated with hypermethylation of DNA (Friedman, 2008). Additionally, when an injury in the liver occurs, growth factors are released in the extracellular space and the extracellular matrix relays signals to cells that trigger a fibrogenic response (Schuppan et al., 2001).

What is notable about this analysis is that, from the perspective of mRNA expression, we expect to find large differences in gene expression between the tissues (Blekhman et al., 2008). Nevertheless, the ANOVA analysis extracts the differentially methylated CpGs associated with these genes, compares their beta values in different tissues and identifies novel relationships between the genes that arise from different methylation patterns.

B.1.3 Correlation analysis

Here, in a similar way to the analysis conducted in Section 4.5.3, we want to determine if there are specific locations where the correlation between methylation and gene expression is over- or under-represented. The two types of locations we considered were: a) location by gene and b) custom location. The main difference between them is the level of granularity. In

TABLE XXXIX

FUNCTIONAL ANNOTATION OF TOP 20 GENES FROM THE ANOVA TEST IN PAI'S STUDY (HUMAN TISSUES).

Tissue comparison	**Id**	**Term**	**Category**	***p*-value**
Kidney vs. Heart	–	None		
Kidney vs. Liver	GO:0002526	Acute inflammatory response	GO BP	1.6e-4
	GO:0048878	Chemical homeostasis	GO BP	1.7e-4
	GO:0009611	Response to wounding	GO BP	2.0e-4
	GO:0005576	Extracellular region	GO CC	1.2e-4
	GO:0004252	Serine-type endopeptidase activity	GO MF	9.9e-4
	GO:0008236	Serine-type peptidase activity	GO MF	1.5e-3
	GO:0017171	Serine hydrolase activity	GO MF	1.5e-3
Heart vs. Liver	GO:0008528	Peptide receptor activity, G-protein coupled	GO MF	2.8e-4
	GO:0042277	Peptide binding	GO MF	1.5e-3

a) we only have two locations: TSS200 and TSS1500 whereas in b) we have TSS200, TSS600, TSS1000 and TSS1500. See Section 4.3.3 for more details.

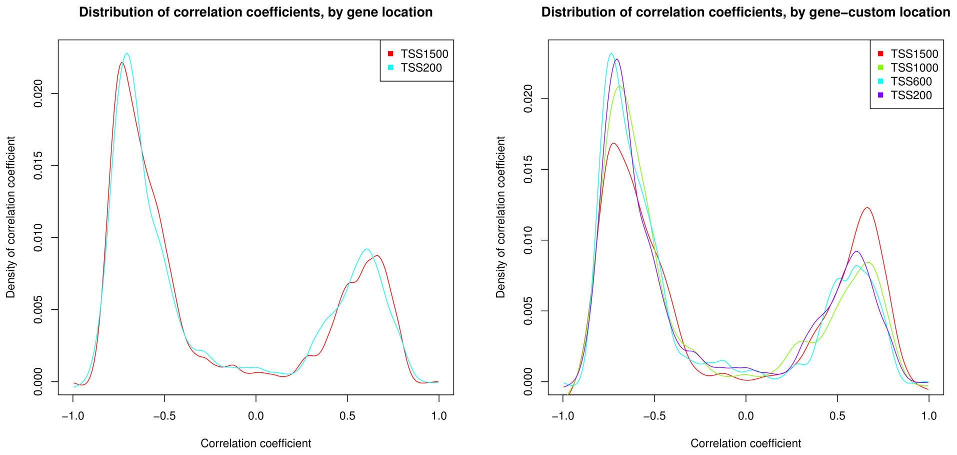
A Fisher's exact test when using only two locations (TSS200 and TS1500) did not identify significant differences between them in any of the three tissue comparisons. In contrast, when testing on the custom locations, different regions showed a significant p-value ($\alpha = 0.01$). Table XL summarizes these results. In general, all custom locations have a two-fold difference between negative and positive correlations. The locations reported by Fisher's exact test contain more correlation coefficients that are negative than positive but at a rate much smaller than two-to-one. In the case of heart vs. liver, Figure 63(a) shows that there are no significant differences between CpGs located less than 200 bp from the TSS (TSS200) and CpGs located up to 1,500 bp from the TSS (TSS1500). If we want to have a more granular look at these locations, we can consider the custom locations in which the TSS1500 group is partitioned into groups with smaller distances. Then we see in Figure 63(b) that TSS1500, the location farthest from the TSS, now has a larger number of positively correlated coefficients. The last row in Table XL shows that this difference is statistically significant.

TABLE XL

FISHER'S EXACT TEST OF CORRELATION COEFFICIENTS AT DIFFERENT GENE LOCATIONS.

Tissue comparison	**Type of location**			
	by gene	***p*-value**	**custom**	***p*-value**
Kidney vs. Heart	–	–	–	–
Kidney vs. Liver	–	–	TSS600	0.004075
			TSS1000	0.000076
Heart vs. Liver	–	–	TSS1500	0.009187

Appendix B (Continued)



(a) Location by gene

(b) Custom location

Figure 63. Density of correlation coefficients for all locations in Heart vs. Liver.

B.1.4 Multiple linear regression

As we did in Section 4.5.4, for each gene we want to find a linear equation to model gene expression with respect to the beta values of the CpGs associated to the gene. The multiple regression analysis yields a p-value per gene and because in the Illumina 27K platform the number of CpGs associated to a gene is much smaller than in the 450K platform, it is more likely that the regression analysis will be statistically significant.

With that in mind, we focused on identifying genes for which a well-fitted linear equation could be found in all tissue comparisons (p-value < 0.01). Our goal was to find differences in methylation patterns between the tissues so we looked at genes with different regression coefficients in the three tissue comparisons (see User Manual for more details about temporary files created by the pipeline that can be programmatically accessed). Among the many genes we identified as having different methylation patterns between the tissues, we found the Potassium voltage-gated channel, Isk-related family, member 3 (KCNE3).

KCNE3 is known to express weakly in the heart (Mazhari et al., 2002). Additionally, the protein encoded by this gene "*...is a type I membrane protein, and a beta subunit that assembles with a potassium channel alpha-subunit to modulate the gating kinetics and enhance stability of the multimeric complex. This gene is prominently expressed in the kidney...*" (Maglott et al., 2005). This can be seen in Table XLI where KCNE3 is always up-regulated in any comparison involving kidney tissue. Table XLI shows the difference in expression levels of KCNE3 and of methylation of its overlapping CpGs. Due to KCNE3's prevalence of expression in the kidney, we will focus only on the two tissue comparisons that involve kidney. Up- or down-regulation refers to the first tissue compared to the second, e.g.: KCNE3 is up-regulated in kidney and, therefore, down-regulated in liver. The same applies for hyper- and hypo-methylation. In the kidney vs. heart comparison, only one CpG is differentially methylated.

We want to address the following question: "Does methylation of KCNE3 in kidney differ substantially from methylation in heart and in liver?" By using the comparison between pairs of tissues as proxy, we can identify different methylation patterns of KCNE3 in heart but not in liver and vice versa.

Figure 64 shows the TSS of KCNE3 (RefSeq Id: NM_005472) obtained from the UCSC Genome Browser. The figure also shows the location of the two CpGs in the proximity of the

Appendix B (Continued)

TABLE XLI

STATUS OF THE GENE KCNE3 AND ITS OVERLAPPING CPGS IN DIFFERENT TISSUES.

Tissue comparison	**Gene status**	**CpG status**	
	KCNE3	**cg02595219**	**cg23189044**
Kidney vs. Liver	up-regulated	hypo-methylated	hypo-methylated
Kidney vs. Heart	up-regulated	hypo-methylated	–

TSS (it is transcribed in the 3'-5' direction). The distances from the CpGs to the TSS are: 223 bp for cg02595219 and 424 bp for cg23189044.

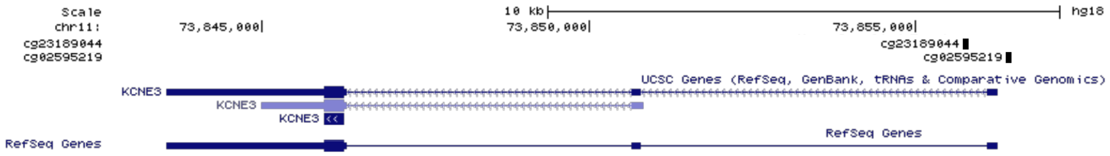


Figure 64. Genomic location of KCNE3 (NM_005472) and its associated CpGs.

Using the model described in Equation 4.3 (Section 4.4.6) we obtained the coefficients and p-values listed in Table XLII. Although the CpGs are located 647 bp apart from each other, their contribution to the linear model seems to be conflicting. In particular, in kidney vs. liver, cg02595219 has a negative coefficient and cg23189044 has a positive one, even when both of them are hypo-methylated. In kidney vs. heart, only cg02595219 is differentially methylated and its negative coefficient is consistent with the kidney vs. liver analysis. This may indicate that lower methylation at that location corresponds to an overall higher expression of the gene. But what about cg23189044 and its positive coefficient?

How can we resolve the apparent contradiction of having positive and negative coefficients? Can we assume that the CpGs with negative coefficients are the ones that have a stronger effect upon gene expression in both tissue comparisons? We can base this hypothesis on the well known fact that hyper-methylation in the promoter of genes is strongly correlated with under-expression of the genes (Stein et al., 1982). But to get a definite answer to our question we need to have a look at the LASSO regression results in the next section.

TABLE XLII

COEFFICIENTS AND P-VALUES OF MULTIPLE LINEAR REGRESSION FOR KCNE3.

		Kidney vs. Liver		**Kidney vs. Heart**	
Coefficients		**Value**	***p*-value**	**Value**	***p*-value**
γ_0	intercept	11.90	2.7e-18	12.99	3.4e-19
γ_1	cg02595219	-5.81	0.1692	-17.63	5.5e-07
γ_2	cg23189044	2.74	0.6635	–	–

B.1.5 LASSO regression

We applied the LASSO method as described in Section 4.4.7 and Equation 4.4 and turned our attention to KCNE3. In the previous section we left unanswered the question of which of

Appendix B (Continued)

the two CpGs in kidney vs. liver has a higher impact on the expression of the gene. The results of LASSO regression shown in Table XLIII help us clarify this issue.

TABLE XLIII

SOLUTION TO LASSO REGRESSION FOR KCNE3.

			Kidney vs. Liver	**Kidney vs. Heart**
	Parameters		**Value**	**Value**
$\hat{\boldsymbol{\alpha}}^k$	α_0	intercept	11.8916	12.9777
	α_1	cg02595219	-3.9515	-17.5099
	α_2	cg23189044	0.0000	–

In the comparison of kidney vs. liver, we see a preference to keep the CpG that was given a negative coefficient by our linear regression analysis (see Table XLII) while cg23189044 is dropped from the model. This is an indication that the stronger effect in KCNE3 comes from the CpG whose beta values have an inverse contribution to the expression level of the gene. We previously suggested this, simply from analyzing the linear regression coefficients. But the LASSO method, by prioritizing the CpGs that best explain the expression values of KCNE3, provides a strong support to our hypothesis.

We repeated the multiple linear regression analysis of kidney vs. liver in KCNE3 excluding cg23189044. There was no need to rerun the regression of kidney vs. heart because it only had one CpG and LASSO preserved it. Table XLIV can be contrasted with Table XLII. When both CpGs were considered, the linear model for KCNE3 attained a statistically significant fit (F-statistic = 13.12, p = 2.6e-4, adjusted R-squared ($\bar{R}^2$) = 0.5359) but after excluding cg23189044, the linear model for KCNE3 has a much better fit and a more significant p-value (F-statistic = 27.14, p = 4.3e-05, $\bar{R}^2$ = 0.5545).

TABLE XLIV

COEFFICIENTS AND P-VALUES OF MULTIPLE LINEAR REGRESSION FOR KCNE3 AFTER FILTERING CPG WITH LASSO.

		Kidney vs.	Liver
Coefficients		**Value**	***p*-value**
γ_0	intercept	11.93	2.47e-19
γ_1	cg02595219	-4.05	4.25e-05

Appendix C

SUPPLEMENTARY INFORMATION FROM CHAPTER 5

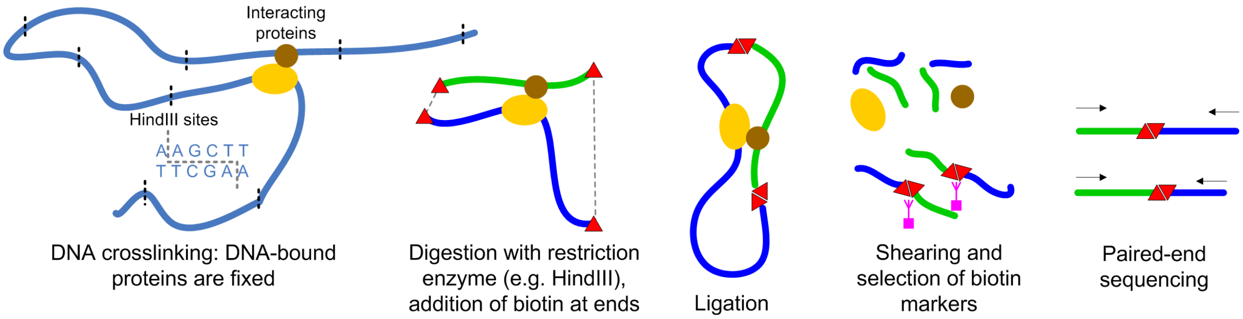

Interacting proteins
HindIII sites
AAGCTT
TTCGAA
DNA crosslinking: DNA-bound proteins are fixed
Digestion with restriction enzyme (e.g. HindIII), addition of biotin at ends
Ligation
Shearing and selection of biotin markers
Paired-end sequencing

Figure 65. An overview of the steps in Hi-C.

Bowtie parameters: Each paired-end was mapped independently of the other. The following syntax is used to map one paired-end. The same command has to be invoked for the other paired-end.

```
bowtie -q -m 1 -k 1 -t --suppress 5,6,7
```

where

Appendix C (Continued)

`-q` : indicates format of input file is `FASTQ`.

`-m 1` : Suppress all alignments for a read if more than 1 reportable alignment exists.

`-k 1` : Report up to 1 valid alignment per read.

`-t` : Verbose mode, print time of each phase.

`-suppress 5,6,7` : Verbose mode, suppress columns 5, 6 and 7 from output. Where

Column 5: Is the read sequence that was mapped/unmapped.

Column 6: Quality scores.

Column 7: A number indicating the number of alignments that were found (valid only when `-m` X with $X > 1$)

Appendix C (Continued)

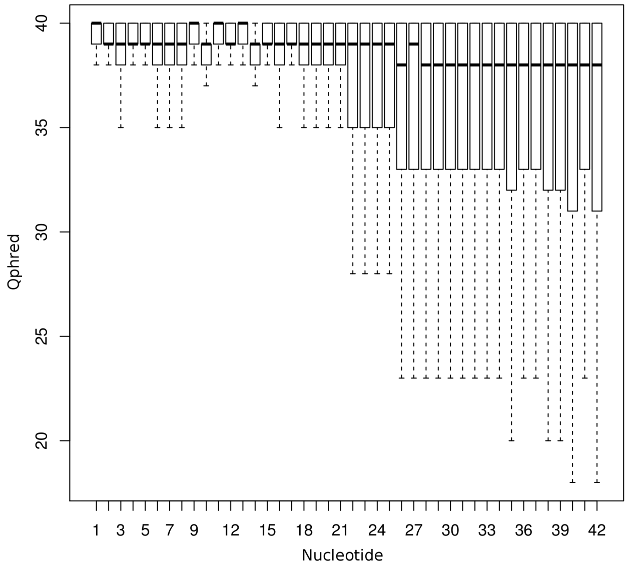


Figure 66. Quality scores of nucleotides (length of read = 42) in reads of MEF replicate 1.

Appendix C (Continued)

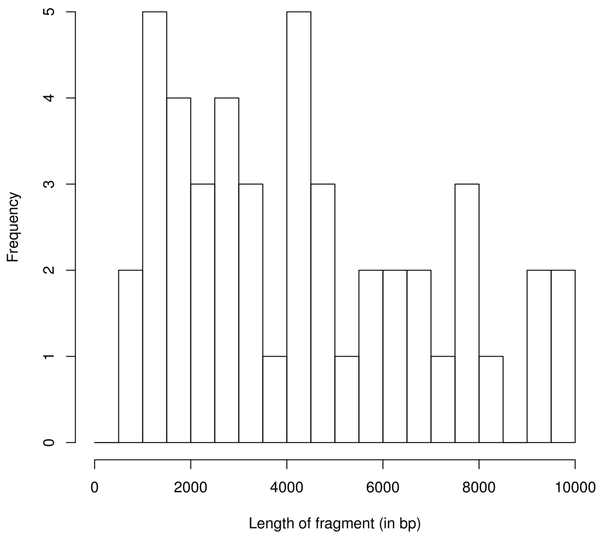


Figure 67. Gene desert in chromosome 5: length of restriction fragments covered by a primer (forward or reverse). Bin size=500 bp.

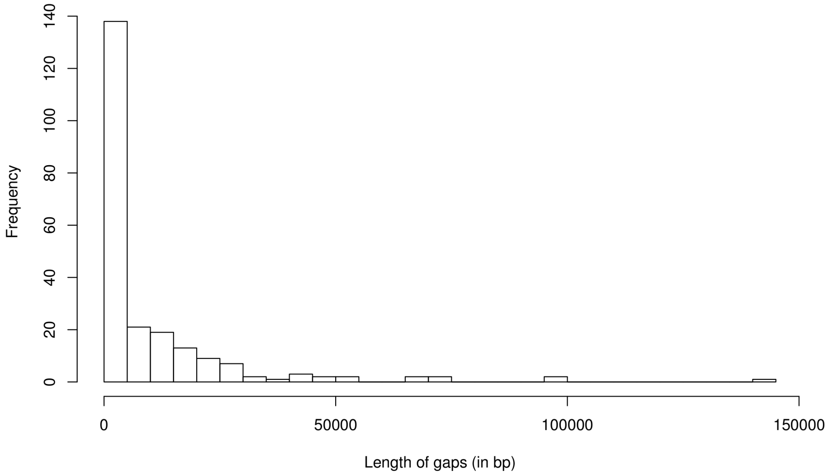


Figure 68. Length of gaps between restriction fragments covered by a primer (forward or reverse) in the Igh locus (chr12). Bin size=5,000 bp.

Appendix C (Continued)

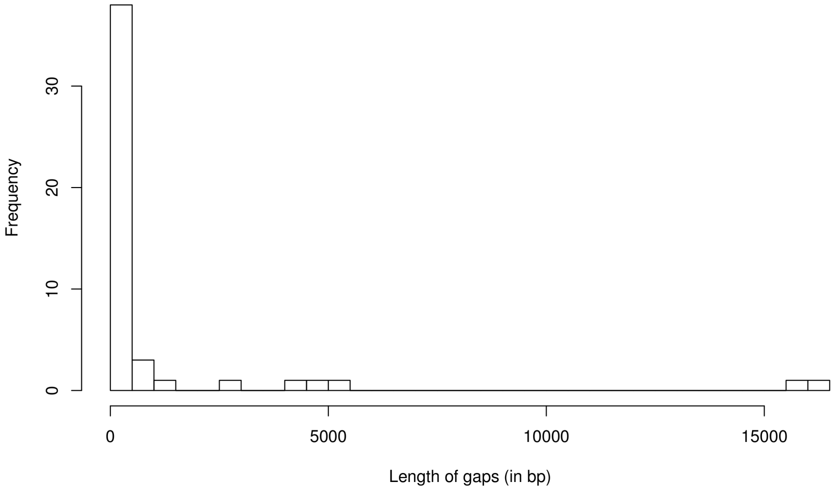


Figure 69. Length of gaps between restriction fragments covered by a primer (forward or reverse) in the gene desert of chr5. Bin size=500 bp.

Appendix C (Continued)

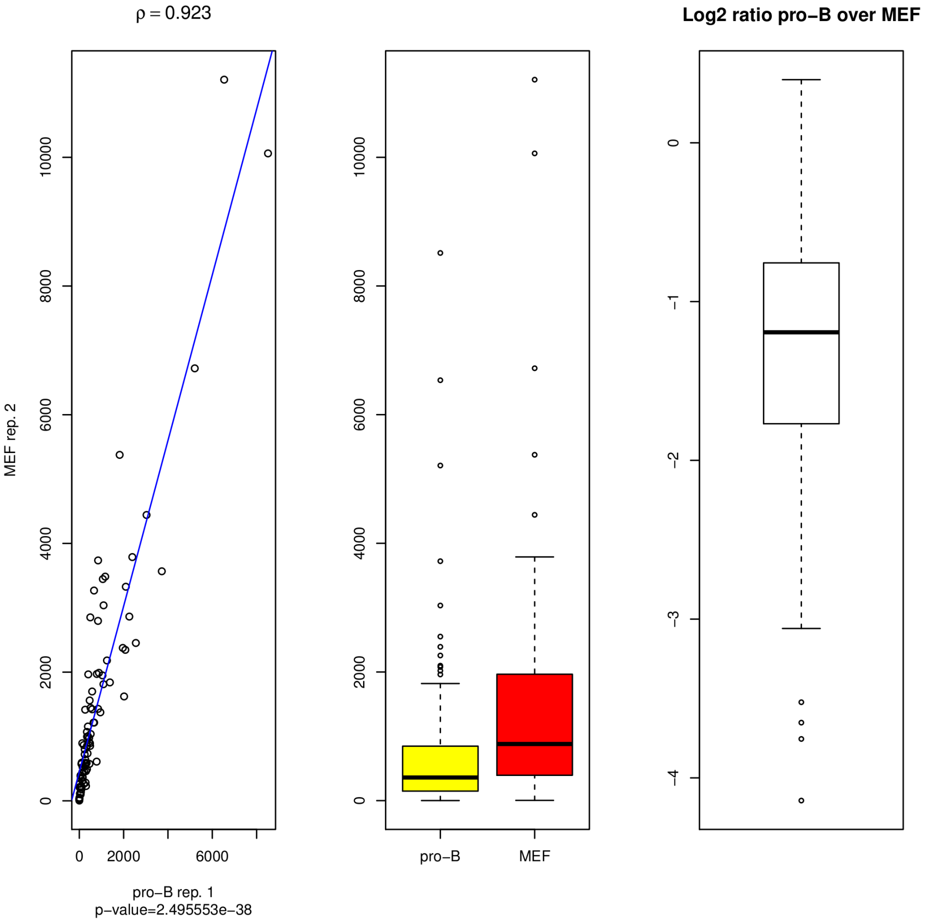


Figure 70. Interactions in chr5, pro-B rep. 1, MEF rep. 2

Appendix C (Continued)

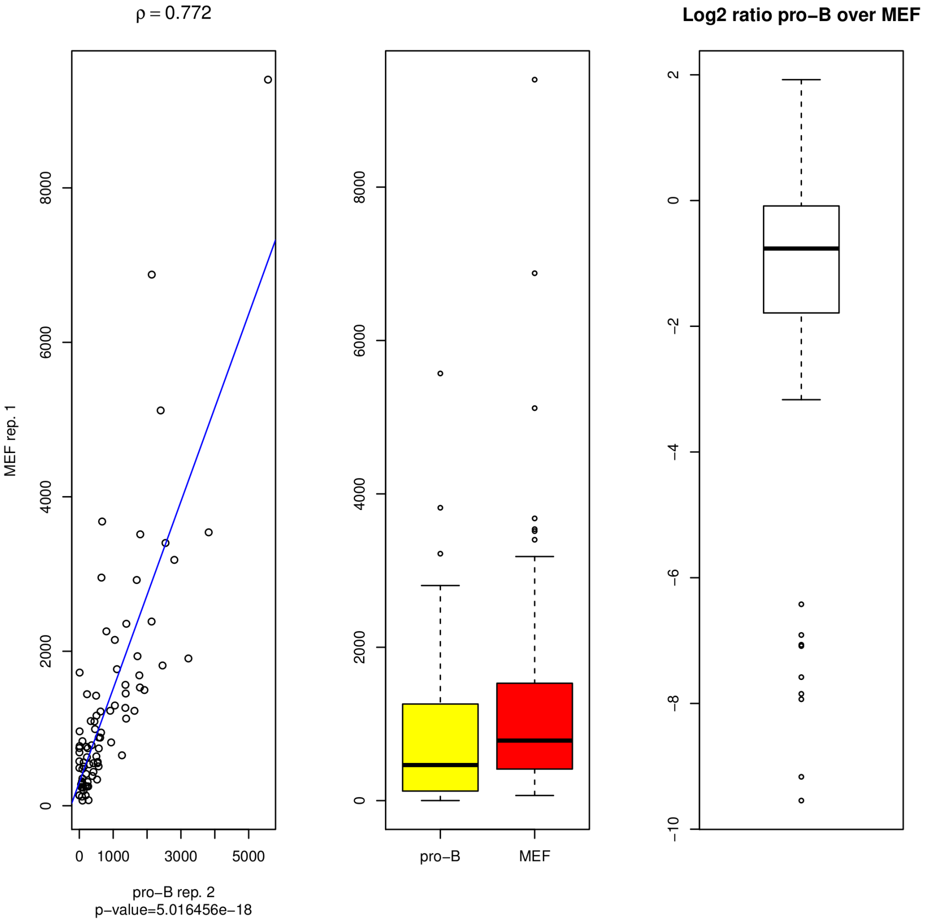


Figure 71. Interactions in chr5, pro-B rep. 2, MEF rep. 1

Appendix C (Continued)

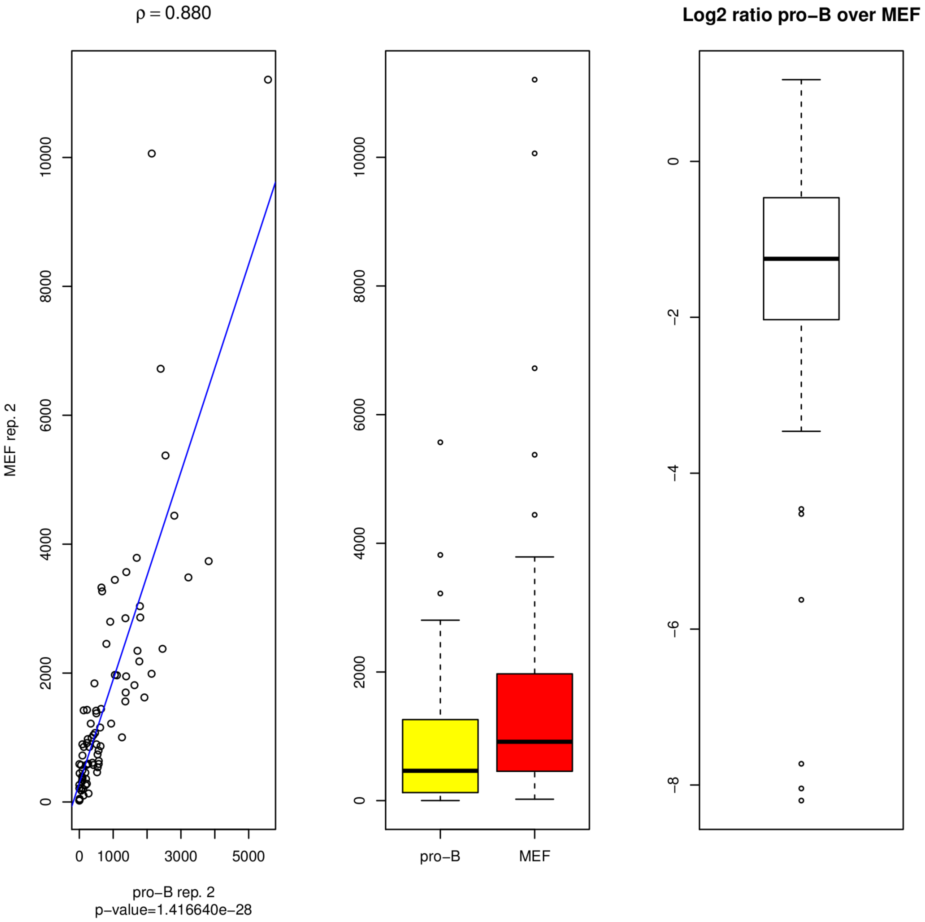


Figure 72. Interactions in chr5, pro-B rep. 2, MEF rep. 2

Appendix C (Continued)

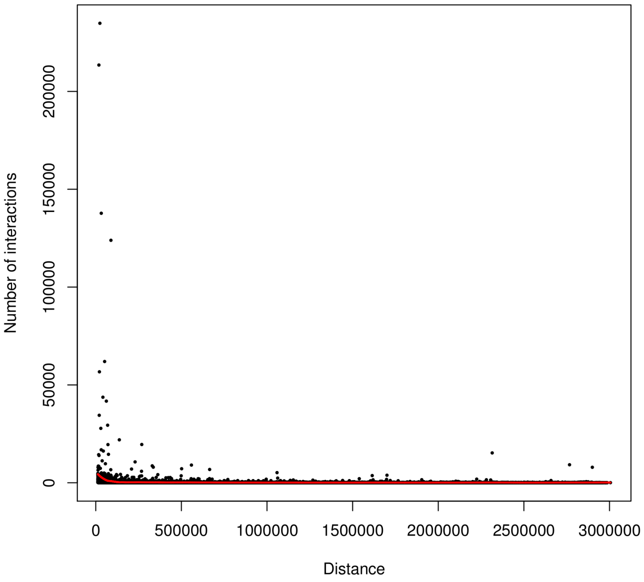


Figure 73. Interactions by distance in MEF with LOESS curve

Appendix C (Continued)

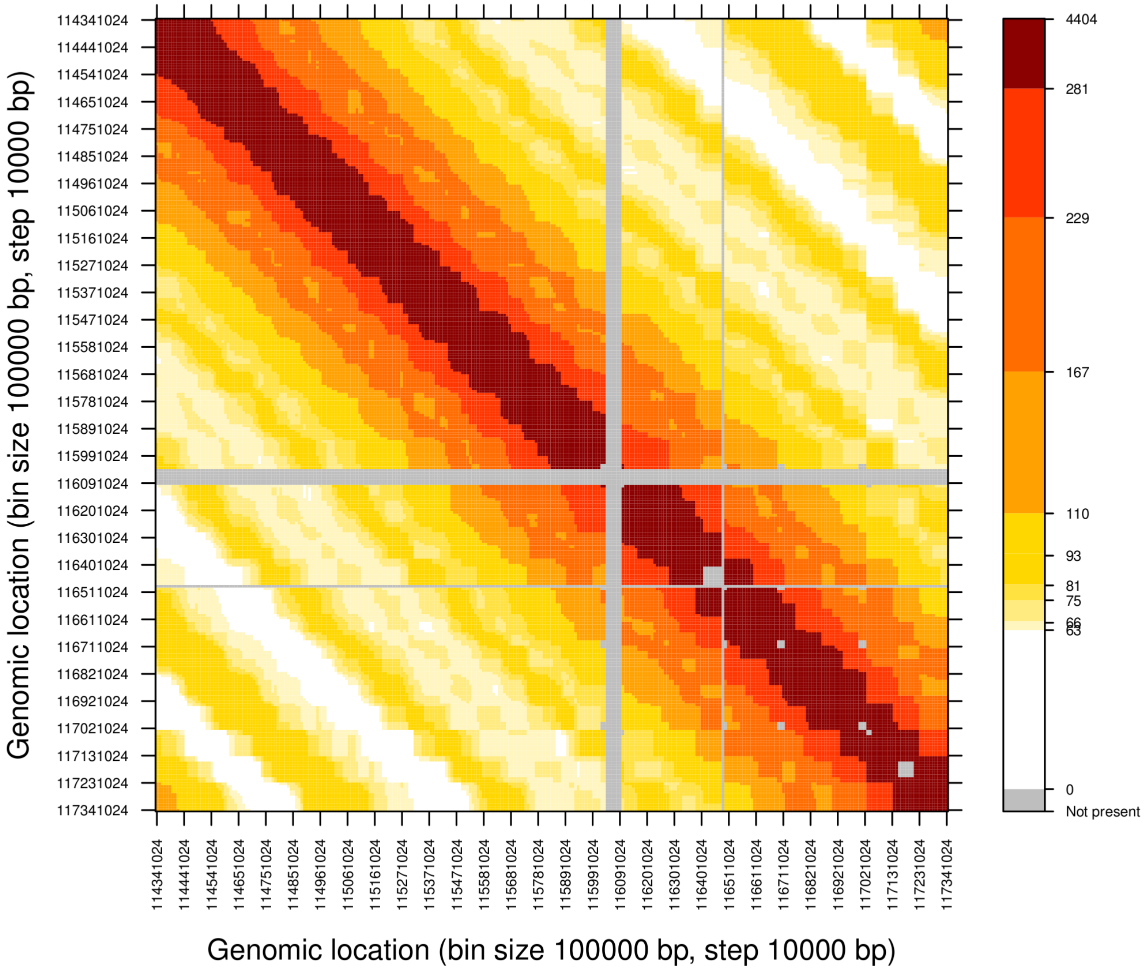


Figure 74. Binned expected interactions in Igh locus in MEF (using LOESS). Bin size=100 Kb, bin step=10 Kb

CITED LITERATURE

Aerts, S., Van Loo, P., Thijs, G., Mayer, H., de Martin, R., Moreau, Y., and De Moor, B.: TOUCAN 2: The all-inclusive open source workbench for regulatory sequence analysis. Nucleic Acids Research, 33(suppl 2):W393–W396, 2005.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P.: Molecular Biology of the Cell, 4th edition.. New York, Garland Science, 4th edition, 2002.

Backes, C., Meese, E., Lenhof, H.-P., and Keller, A.: A dictionary on microRNAs and their putative target pathways. Nucleic Acids Research, 2010.

Bailey, T. and Elkan, C.: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pages 28–36. AAAI Press, 1994.

Basham, B., Sathe, M., Grein, J., McClanahan, T., DAndrea, A., Lees, E., and Rascle, A.: In vivo identification of novel stat5 target genes. Nucleic Acids Research, 36(11):3802–3818, 2008.

Baù, D., Sanyal, A., Lajoie, B. R., Capriotti, E., Byron, M., Lawrence, J. B., Dekker, J., and Marti-Renom, M. A.: The three-dimensional folding of the -globin gene domain reveals formation of chromatin globules. Nat Struct Mol Biol, 18(1):107–114, January 2011.

Beckmann, H., Su, L. K., and Kadesch, T.: TFE3: a helix-loop-helix protein that activates transcription through the immunoglobulin enhancer muE3 motif. Genes & Development, 4(2):167–179, 1990.

Bejerano, G.: Efficient exact value computation and applications to biosequence analysis. In Proceedings of the seventh annual international conference on Research in computational molecular biology, RECOMB ’03, pages 38–47, New York, NY, USA, 2003. ACM.

Benjamini, Y. and Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B Methodological, 57:289–300, 1995.

Bibikova, M., Lin, Z., Zhou, L., Chudin, E., Garcia, E. W., Wu, B., Doucet, D., Thomas, N. J., Wang, Y., Vollmer, E., Goldmann, T., Seifart, C., Jiang, W., Barker, D. L., Chee, M. S., Floros, J., and Fan, J.-B.: High-throughput dna methylation profiling using universal bead arrays. Genome Research, 16(3):383–393, 2006.

Blekhman, R., Oshlack, A., Chabot, A. E., Smyth, G. K., and Gilad, Y.: Gene regulation in primates evolves under tissue-specific selection pressures. PLoS Genet, 4(11):e1000271, 11 2008.

Boersma, B. J., Reimers, M., Yi, M., Ludwig, J. A., Luke, B. T., Stephens, R. M., Yfantis, H. G., Lee, D. H., Weinstein, J. N., and Ambs, S.: A stromal gene signature associated with inflammatory breast cancer. International Journal of Cancer, 122(6):1324–1332, 2008.

Breiman, L.: Random Forests. Machine Learning, 45:5–32, 2001. 10.1023/A:1010933404324.

Brönneke, S., Brckner, B., Peters, N., Bosch, T. C., Stäb, F., Wenck, H., Hagemann, S., and Winnefeld, M.: DNA methylation regulates lineage-specifying genes in primary lymphatic and blood endothelial cells. Angiogenesis, 15(2):317–329, 2012.

Bryne, J. C., Valen, E., Tang, M.-H. E., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B., and Sandelin, A.: JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Research, 36(suppl 1):D102–D106, 2008.

Buffa, F. M., Camps, C., Winchester, L., Snell, C. E., Gee, H. E., Sheldon, H., Taylor, M., Harris, A. L., and Ragoussis, J.: microRNA-Associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. Cancer Research, 71(17):5635–5645, 2011.

Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M., and Werner, T.: MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics, 21(13):2933–2942, 2005.

Chang, L.-W., Fontaine, B. R., Stormo, G. D., and Nagarajan, R.: PAP: A comprehensive workbench for mammalian transcriptional regulatory sequence analysis. Nucleic Acids Research, 35(suppl 2):W238–W244, 2007.

Cheng, C. and Gerstein, M.: Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. Nucleic Acids Research, 2011.

Cheng, J. and Druzdzel, M. J.: AIS-BN: An adaptive importance sampling algorithm for evidential reasoning in large Bayesian networks. Journal of Artificial Intelligence Research, 13:13–155, 2000.

Cleveland, W. S.: Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association, 74(368):829–836, December 1979.

Cleveland, W. S. and Devlin, S. J.: Locally weighted regression: An approach to regression analysis by local fitting. Journal of the American Statistical Association, 83(403):596–610, September 1988.

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M.: The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research, 38(6):1767–1771, 2010.

Cole, S. W., Yan, W., Galic, Z., Arevalo, J., and Zack, J. A.: Expression-based monitoring of transcription factor activity: the TELiS database. Bioinformatics, 21(6):803–810, 2005.

Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E.: WebLogo: A sequence logo generator. Genome Research, 14(6):1188–1190, 2004.

da Fonseca, P. G. S., Guimarães, K. S., and Sagot, M.-F.: Efficient representation and P-value computation for high-order Markov motifs. Bioinformatics, 24(16):i160–i166, 2008.

Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., Bunney, W. E., Myers, R. M., Speed, T. P., Akil, H., Watson, S. J., and Meng, F.: Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Research, 33(20):e175, 2005.

Davies, L., Spiller, D., White, M. R. H., Grierson, I., and Paraoan, L.: PERP expression stabilizes active p53 via modulation of p53-MDM2 interaction in uveal melanoma cells. Cell Death and Dis, 2:e136, March 2011.

Deaton, A. M. and Bird, A.: CpG islands and the regulation of transcription. Genes & Development, 25(10):1010–1022, 2011.

Dekker, J.: The three ‘C’ s of chromosome conformation capture: controls, controls, controls. Nat Meth, 3(1):17–21, January 2006.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N.: Capturing chromosome conformation. Science, 295(5558):1306–1311, 2002.

Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L.: Improved microbial gene identification with GLIMMER. Nucleic Acids Research, 27(23):4636–4641, 1999.

Desmedt, C., Haibe-Kains, B., Wirapati, P., Buyse, M., Larsimont, D., Bontempi, G., Delorenzi, M., Piccart, M., and Sotiriou, C.: Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. Clinical Cancer Research, 14(16):5158–5165, 2008.

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., Green, R. D., and Dekker, J.: Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. Genome Research, 16(10):1299–1309, 2006.

Du, P., Kibbe, W. A., and Lin, S. M.: lumi: a pipeline for processing Illumina microarray. Bioinformatics, 24(13):1547–1548, 2008.

Edgar, R., Domrachev, M., and Lash, A. E.: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research, 30(1):207–210, 2002.

ENCODE, P. C.: An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414):57–74, September 2012.

Enerly, E., Steinfeld, I., Kleivi, K., Leivonen, S.-K., Aure, M. R., Russnes, H. G., Rnneberg, J. A., Johnsen, H., Navon, R., Rdland, E., Mkel, R., Naume, B., Perl, M., Kallioniemi, O., Kristensen, V. N., Yakhini, Z., and Brresen-Dale, A.-L.: miRNA-mRNA Integrated analysis reveals roles for miRNAs in primary breast tumors. PLoS ONE, 6(2):e16915, 02 2011.

Fraser, J., Ethier, S. D., Miura, H., and Dostie, J.: Chapter five - a torrent of data: Mapping chromatin organization using 5C and high-throughput sequencing. In Nucleosomes, Histones and Chromatin Part B, eds. C. Wu and C. D. Allis, volume 513 of Methods in Enzymology, pages 113 – 141. Academic Press, 2012.

Friedman, J., Hastie, T., and Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J Stat Softw, 33(1):1–22, 2010.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D.: Using Bayesian networks to analyze expression data. Journal of Computational Biology, 7(5):601–20, 2000.

Friedman, R. C., Farh, K. K.-H., Burge, C. B., and Bartel, D. P.: Most mammalian mRNAs are conserved targets of microRNAs. Genome Research, 19(1):92–105, 2009.

Friedman, S. L.: Mechanisms of hepatic fibrogenesis. Gastroenterology, 134:1655–1669, 2008.

Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., Goldman, M., Barber, G. P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T. R., Giardine, B. M., Harte, R. A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R. M., Learned, K., Li, C. H., Meyer, L. R., Pohl, A., Raney, B. J., Rosenbloom, K. R., Smith, K. E., Haussler, D., and Kent, W. J.: The UCSC Genome Browser database: update 2011. Nucleic Acids Research, 2010.

Geller, S. C., Gregg, J. P., Hagerman, P., and Rocke, D. M.: Transformation and normalization of oligonucleotide microarray data. Bioinformatics, 19(14):1817–1823, 2003.

Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J., and Zhang, J.: Bioconductor: open software development for computational biology and bioinformatics. Genome Biology, 5(10):R80, 2004.

Greally, J. M.: Bidding the CpG island goodbye. eLife, 2, 2013.

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics). Springer, 2009.

Ho Sui, S. J., Fulton, D. L., Arenillas, D. J., Kwon, A. T., and Wasserman, W. W.: oPOSSUM: Integrated tools for analysis of regulatory motif over-representation. Nucleic Acids Research, 35(suppl 2):W245–W252, 2007.

Hua, X., Miller, Z. A., Benchabane, H., Wrana, J. L., and Lodish, H. F.: Synergism between Transcription Factors TFE3 and Smad3 in Transforming Growth Factor-β-induced Transcription of the Smad7 Gene. Journal of Biological Chemistry, 275(43):33205–33208, 2000.

Hua, X., Miller, Z. A., Wu, G., Shi, Y., and Lodish, H. F.: Specificity in transforming growth factor β-induced transcription of the plasminogen activator inhibitor-1 gene: Interactions of promoter DNA, transcription factor μE3, and Smad proteins. Proceedings of the National Academy of Sciences, 96(23):13130–13135, 1999.

Huang, D. W., Sherman, B. T., and Lempicki, R. A.: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protocols, 4(1):44–57, December 2008.

Huang, G. T., Athanassiou, C., and Benos, P. V.: mirConnX: condition-specific mRNA-microRNA network integrator. Nucleic Acids Research, 39(suppl 2):W416–W423, 2011.

Huang, L., Roqueiro, D., and Dai, Y.: Analyzing mRNA and microRNA co-expression profiles to identify pathways and their potential regulators in ER+ and ER- breast tumors. In Engineering in Medicine and Biology Society,EMBC, 2011 Annual International Conference of the IEEE, pages 932 –935, 30 2011-sept. 3 2011.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., de Castro, E., Coggill, P., Corbett, M., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Fraser, M., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., McMenamin, C., Mi, H., Mutowo-Muellenet, P., Mulder, N., Natale, D., Orengo, C., Pesseat, S., Punta, M., Quinn, A. F., Rivoire, C., Sangrador-Vegas, A., Selengut, J. D., Sigrist, C. J. A., Scheremetjew, M., Tate, J., Thimmajanarthanan, M., Thomas, P. D., Wu, C. H., Yeats, C., and Yong, S.-Y.: InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Research, 40(D1):D306–D312, 2012.

Hyndman, R. J. and Fan, Y.: Sample quantiles in statistical packages. The American Statistician, 50(4):361–365, 1996.

Illumina: Infinium® Humanmethylation27 BeadChip. Data sheet: Epigenetics. 2010.

Illumina: Infinium® Humanmethylation450 BeadChip. Data sheet: Epigenetics. 2012.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P.: Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Research, 31(4):e15, 2003.

Jensen, F. V. and Nielsen, T. D.: Bayesian Networks and Decision Graphs, Second Edition.. Springer Verlag, 2007.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B.: Genome-wide mapping of in vivo protein-DNA interactions. Science, 316(5830):1497–1502, 2007.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M.: KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Research, 38(suppl 1):D355–D360, 2010.

Kel, A., Gößling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O., and Wingender, E.: MATCH^{TM}: a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Research, 31(13):3576–3579, 2003.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., Haussler, and David: The human genome browser at UCSC. Genome Research, 12(6):996–1006, 2002.

Krek, A., Grun, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M., and Rajewsky, N.: Combinatorial microRNA target predictions. Nat Genet, 37(5):495–500, May 2005.

Kwisthout, J.: Most probable explanations in Bayesian networks: Complexity and tractability. Int. J. Approx. Reasoning, 52(9):1452–1469, December 2011.

Lajoie, B. R., van Berkum, N. L., Sanyal, A., and Dekker, J.: My5C: web tools for chromosome conformation capture studies. Nat Meth, 6(10):690–691, October 2009.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10(3):R25, 2009.

Le Bechec, A., Portales-Casamar, E., Vetter, G., Moes, M., Zindy, P.-J., Saumet, A., Arenillas, D., Theillet, C., Wasserman, W., Lecellier, C.-H., and Friederich, E.: MIR@NT@N: a framework integrating transcription factors, microRNAs and their targets to identify subnetwork motifs in a meta-regulation network model. BMC Bioinformatics, 12(1):67, 2011.

Lewin, B., Krebs, J. E., Kilpatrick, S. T., and Goldstein, E. S.: Lewin's genes X. Sudbury, Mass., Jones and Bartlett, 2011.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander,

E. S., and Dekker, J.: Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 326(5950):289–293, 2009.

Liu, X., Brutlag, D. L., and Liu, J. S.: Bioprospector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In Pac. Symp. Biocomput, pages 127–138, 2001.

Longley, D. B. and Johnston, P. G.: 5-Fluorouracil. Apoptosis, Cell Signaling, and Human Diseases.. Humana Press Inc., 2007.

Lopez-Romero, P.: Pre-processing and differential expression analysis of Agilent microRNA arrays using the AgiMicroRna Bioconductor library. BMC Genomics, 12(1):64, 2011.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T.: Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research, 33(suppl 1):D54–D58, 2005.

Maksimovic, J., Gordon, L., and Oshlack, A.: SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. Genome Biology, 13(6):R44, 2012.

Mardis, E. R.: Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics, 9(1):387–402, 2008. PMID: 18576944.

Martinez, N. J. and Walhout, A. J. M.: The interplay between transcription factors and microRNAs in genome-scale regulatory networks. BioEssays, 31(4):435–445, 2009.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., and Wingender, E.: TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. Nucleic Acids Research, 34(suppl 1):D108–D110, 2006.

Mazhari, R., Nuss, H. B., Armoundas, A. A., Winslow, R. L., and Marbn, E.: Ectopic expression of KCNE3 accelerates cardiac repolarization and abbreviates the QT interval. The Journal of Clinical Investigation, 109(8):1083–1090, 4 2002.

Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E. T., and Bergh, J.: An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proceedings of the National Academy of Sciences of the United States of America, 102(38):13550–13555, 2005.

Minn, A. J., Gupta, G. P., Siegel, P. M., Bos, P. D., Shu, W., Giri, D. D., Viale, A., Olshen, A. B., Gerald, W. L., and Massague, J.: Genes that mediate breast cancer metastasis to lung. Nature, 436(7050):518–524, July 2005.

Murphy, K.: Janeway's Immunobiology. New York, Garland Science, Taylor & Francis Group, 8th edition, July 2011.

Murphy, K.: The Bayes Net Toolbox for Matlab. Computing Science and Statistics: Proceedings of Interface, 33, 2001.

Osborne, C. K. and Schiff, R.: Mechanisms of endocrine resistance in breast cancer. Annual Review of Medicine, 62(1):233–247, 2011. PMID: 20887199.

Pai, A. A., Bell, J. T., Marioni, J. C., Pritchard, J. K., and Gilad, Y.: A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. PLoS Genet, 7(2):e1001316, 02 2011.

Pearl, J.: Bayesian networks: A model of self-activated memory for evidential reasoning. In Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine, pages 329–334, August 1985.

Phillips-Cremins, J., Sauria, M., Sanyal, A., Gerasimova, T., Lajoie, B., Bell, J., Ong, C.-T., Hookway, T., Guo, C., Sun, Y., Bland, M., Wagstaff, W., Dalton, S., McDevitt, T., Sen, R., Dekker, J., Taylor, J., and Corces, V.: Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell, 153(6):1281–1295, 2013.

Pollard, S. M., Stricker, S. H., and Beck, S.: A shore sign of reprogramming. Cell Stem Cell, 5(6):571–572, 2009.

Quackenbush, J.: Microarray data normalization and transformation. Nat Genet, 2002.

Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T.: Matlnd and Matlnspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Research, 23(23):4878–4884, 1995.

Roqueiro, D., Frasor, J., and Dai, Y.: BindSDb: A binding-information spatial database. In Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference on, pages 573 –578, 2010.

Rouaud, P., Vincent-Fabert, C., Saintamand, A., Fiancette, R., Marquet, M., Robert, I., Reina-San-Martin, B., Pinaud, E., Cogné, M., and Denizot, Y.: The IgH 3' regulatory region controls somatic hypermutation in germinal center B cells. The Journal of Experimental Medicine, 210(8):1501–1507, 2013.

Sanyal, A., Lajoie, B. R., Jain, G., and Dekker, J.: The long-range interaction landscape of gene promoters. Nature, 489(7414):109–113, September 2012.

Saxonov, S., Berg, P., and Brutlag, D. L.: A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proceedings of the National Academy of Sciences of the United States of America, 103(5):1412–1417, 2006.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. H.: PID: the Pathway Interaction Database. Nucleic Acids Research, 37(suppl 1):D674–D679, 2009.

Schneider, T. D. and Stephens, R.: Sequence logos: a new way to display consensus sequences. Nucleic Acids Research, 18(20):6097–6100, 1990.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P.: SMART, a simple modular architecture research tool: Identification of signaling domains. Proceedings of the National Academy of Sciences, 95(11):5857–5864, 1998.

Schuppan, D., Ruehl, M., Somasundaram, R., and Hahn, E. G.: Matrix as a modulator of hepatic fibrogenesis. Semin Liver Dis, 21(03):351–372, 2001.

Shendure, J. and Ji, H.: Next-generation DNA sequencing. Nat Biotech, 26(10):1135–1145, October 2008.

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

Smyth, G. K. and Speed, T.: Normalization of cDNA microarray data. Methods, 31(4):265–273, December 2003. Candidate Genes from DNA Array Screens: application to neuroscience.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., Van de Vijver, M. J., Bergh, J., Piccart, M., and Delorenzi, M.: Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis. Journal of the National Cancer Institute, 98(4):262–272, 2006.

Stein, R., Razin, A., and Cedar, H.: In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. Proceedings of the National Academy of Sciences, 79(11):3418–3422, 1982.

Sun, H. and Wang, S.: Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. Bioinformatics, 28(10):1368–1375, 2012.

Suzuki, M. M. and Bird, A.: DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet, 9(6):465–476, June 2008.

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouz, P., and Moreau, Y.: A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics, 17(12):1113–1122, 2001.

Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouzé, P., and Moreau, Y.: A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. Journal of Computational Biology, 9(suppl 2):447–464, 2004.

Tibshirani, R.: Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(3):273–282, 2011.

Touzet, H. and Varré, J.-S.: Efficient and accurate p-value computation for position weight matrices. Algorithms for Molecular Biology, 2(1):15, 2007.

Veerla, S., Ringner, M., and Hoglund, M.: Genome-wide transcription factor binding site/promoter databases for the analysis of gene sets and co-occurrence of transcription factor binding motifs. BMC Genomics, 11(1):145, 2010.

Wang, D., Yan, L., Hu, Q., Sucheston, L. E., Higgins, M. J., Ambrosone, C. B., Johnson, C. S., Smiraglia, D. J., and Liu, S.: Ima: An r package for high-throughput analysis of illuminas 450k infinium methylation data. Bioinformatics, 2012.

Wang, K. C., Yang, Y. W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B. R., Protacio, A., Flynn, R. A., Gupta, R. A., Wysocka, J., Lei, M., Dekker, J., Helms,

J. A., and Chang, H. Y.: A long noncoding rna maintains active chromatin to coordinate homeotic gene expression. Nature, 472(7341):120–124, April 2011.

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., van Gelder, M. E. M., Yu, J., Jatkoe, T., Berns, E. M., Atkins, D., and Foekens, J. A.: Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet, 365(9460):671–679, 2005.

Warden, C. D., Lee, H., Tompkins, J. D., Li, X., Wang, C., Riggs, A. D., Yu, H., Jove, R., and Yuan, Y.-C.: Cohcap: an integrative genomic pipeline for single-nucleotide resolution dna methylation analysis. Nucleic Acids Research, 41(11):e117, 2013.

Wasserman, W. W. and Sandelin, A.: Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet, 5(4):276–287, April 2004.

West, A. G. and Fraser, P.: Remote control of gene transcription. Human Molecular Genetics, 14(suppl 1):R101–R111, 2005.

Wingender, E., Dietze, P., Karas, H., and Knppel, R.: TRANSFAC: A Database on Transcription Factors and Their DNA Binding Sites. Nucleic Acids Research, 24(1):238–241, 1996.

Yaffe, E. and Tanay, A.: Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet, 43(11), November 2011.

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

Zhang, J. D. and Wiemann, S.: KEGGgraph: a graph approach to KEGG PATHWAY in R and Bioconductor. Bioinformatics, 25(11):1470–1471, 2009.

Zhao, Z., Tavoosidana, G., Sjolinder, M., Gondor, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Singh Sandhu, K., Singh, U., Pant, V., Tiwari, V., Kurukuti, S., and Ohlsson, R.: Circular Chromosome Conformation Capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet, 38(11):1341–1347, November 2006.

VITA

Education

08/2007–present
PhD candidate, Bioinformatics. University of Illinois at Chicago (UIC).
Chicago, Illinois, USA.
Cumulative GPA: 3.83 / 4.0

07/2003
Master of Science in Computer Science. University of Illinois at Chicago.
Cumulative GPA: 3.87 / 4.0

03/1997
Information Systems Engineer. Universidad Tecnológica Nacional.
Buenos Aires, Argentina.
Cumulative GPA: 8.33 / 10 (best GPA of my class)

Academic experience

Teaching assistant
08/2012–05/2013
College of Engineering, Bioinformatics online courses.
- BIOE 582 Computational Functional Genomics, Spring 2013.
- BIOE 439 Biostatistics, Fall 2012.

Research assistant
07/2010–06/2012
Laboratory of Computational Functional Genomics, UIC.
Conducted research with my academic advisor, Dr. Yang Dai.

Research assistant
06/2009–07/2010
Laboratory for Advanced Computing, UIC.
- Implementation of peak callers (for Affymetrix, Agilent and Solexa/Illumina platforms). These tools were used in the context of cloud computing by the Institute of Genomics and Systems Biology (IGSB).
- Design and implementation of a system to launch analysis pipelines from a web interface. This system was implemented in the Bionimbus cloud.

Graduate assistant
05/2007–05/2009
Graduate College, UIC.
- Design and implementation of administrative web-based applications using Oracle Application Express in Oracle 10g database.
- General IT support.

Research assistant
08/2003–09/2005
Midwest Latino Health Research, Training and Policy Center at UIC .
- Design, creation and maintenance of the center's website.
- Creation of analytical reports for grant proposal using SPSS and SAS.

Professional experience

10/2005–05/2007 **Accenture Technology Labs**, Chicago.

- Researcher in the automatic surveillance and sensor fusion project. This work attempted to track people in an indoor environment monitored by video cameras and other sensors.

12/1996–07/2001 **Electronic Data Systems (EDS)** – Buenos Aires, Argentina (currently Hewlett-Packard).

Software Developer, Banco Rio Account.

- Analysis and design of process specifications. Programming in Visual C++.
- SQL programming (SQL Server 6.X) of stored procedures and triggers. Query optimization and administration of the application databases.

Systems & Security Auditor, Interbanking Account.

- Programming in Visual C++ and Visual Basic to implement revisions on system logs and automatic e-mailing (MAPI programming) of notifications.
- OS/390 Mainframe: Change control of the operating system parameters; Security policies and life cycle of programs in the production environment.

Tivoli Engineer, General Motors (GM) & GMAC Operations Account.

- Technical leader in the design, implementation and deployment of Tivoli Framework (3.2 & 3.6.2) for GM and GMAC.

02/1996–09/1996 **Olivetti Argentina** – Buenos Aires, Argentina.

- Development of drivers and interfaces in Visual C++ for proprietary peripheral devices.

Lecturer
03/1997–08/2001 **Universidad Tecnológica Nacional** – Buenos Aires, Argentina.
(National University of Technology).

- Lecturer in Introduction to Programming courses (for freshmen).

Academic involvement

Conference reviewer — (EMBC) Engineering in Medicine and Biology Conference of the IEEE.

03/2013 • EMBC'13.
05/2011 • EMBC'11.
06/2010 • EMBC'10.

Journal reviewer — Journal of Computer Science and Technology.

01/2012 • Volume 27

Honors & awards

08/2012–07/2017 **UIC–Chancellor's Graduate Research Fellowship** to conduct research in the area of chromosome conformation capture. Primary investigator: Dr. Amy Kenter, Dept. of Microbiology and Immunolgy, College of Medicine, UIC.

11/2011 "IEEE BIBM'11 Conference, PhD Student Author travel award"

04/2008 "UIC–Chancellor's Student Service Award" for volunteer service at the Emergency room, UIC Hospital.

03/2003 "UIC–Chancellor's Student Service Award" for service to the community.

08/2001-07/2003 **Fulbright Fellowship**, Masters Program 2001.

Volunteer activities

09/2005–11/2008 Little Brothers Friends of the Elderly, Visiting volunteer.

11/2006–12/2010 UIC Medical Center, Volunteer in the Emergency Department.

Publications in Bioinformatics

Journals

2013 Yalda Afshar, Julie Hastings, Damian Roqueiro, Jae-Wook Jeong, Linda C. Giudice, and Asgerally T. Fazleabas. Changes in eutopic endometrial gene expression during the progression of experimental endometriosis in the baboon, papio anubis. Biology of Reproduction, 2013

2013 Wenbo Mu, Damian Roqueiro, and Yang Dai. A local genetic algorithm for the identification of condition-specific microrna-gene modules. The Scientific World Journal, 2013:9, 2013

2012 Damian Roqueiro, Lei Huang, and Yang Dai. Identifying transcription factors and microRNAs as key regulators of pathways using Bayesian inference on known pathway structures. Proteome Sci, 10 Suppl 1:S15, 2012

2012 Ping Yin, Damian Roqueiro, Lei Huang, Jonas K. Owen, Anna Xie, Antonia Navarro, Diana Monsivais, John S. Coon V, J. Julie Kim, Yang Dai, and Serdar E. Bulun. Genome-wide progesterone receptor binding: Cell type-specific and shared mechanisms in T47D breast cancer cells and primary leiomyoma cells. PLoS ONE, 7(1):e29021, 01 2012

2012 Yalda Afshar, Jae-Wook Jeong, Damian Roqueiro, Franco DeMayo, John Lydon, Freddy Radtke, Rachel Radnor, Lucio Miele, and Asgerally Fazleabas. Notch1 mediates uterine stromal differentiation and is critical for complete decidualization in the mouse. The FASEB Journal, 26(1):282–294, 2012

Conferences

2011 Damian Roqueiro, Lei Huang, and Yang Dai. Identifying transcription factors and microRNAs as key regulators of pathways using Bayesian inference on known pathway structures. In Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference, pages 228–233, 2011

2011 Lei Huang, Damian Roqueiro, and Yang Dai. Analyzing mRNA and microRNA co-expression profiles to identify pathways and their potential regulators in ER+ and ER- breast tumors. In Engineering in Medicine and Biology Society (EMBC), 2011 Annual International Conference of the IEEE, pages 932–935, 2011

2011 Hong Hu, Damian Roqueiro, and Yang Dai. Prioritizing predicted cis-regulatory elements for co-expressed gene sets based on lasso regression models. In Engineering in Medicine and Biology Society,EMBC, 2011 Annual International Conference of the IEEE, pages 6853–6856, 2011

Workshops

2010 Damian Roqueiro, Jonna Frasor, and Yang Dai. BindSDb: A binding-information spatial database. In Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference, pages 573 –578, 2010

Publications in Computer Science

Journals

2007 Damian Roqueiro and Valery A. Petrushin. Counting people using video cameras. International Journal of Parallel, Emergent and Distributed Systems, 22(3):193–209, 2007

Conferences

2006 Valery A. Petrushin, Omer Shakil, Damian Roqueiro, Gang Wei, and Anatole V. Gershman. Multiple-sensor indoor surveillance system. In Proceedings of The 3rd Canadian Conference on Computer and Robot Vision (CRV'06), 2006

Workshops

2007 Anatole Gershman, Rayid Ghani, Damian Roqueiro, and Gang Wei. Trade-offs in the use of Bayesian filtering for sensor fusion. In International Workshop on Knowledge Discovery from Sensor Data (Sensor-KDD'07), 2007

2006 Damian Roqueiro and Valery A. Petrushin. Counting people using video cameras. In Workshop for Multimedia and Data Mining (MDM at KDD'06), 2006

2006 Bradley P. Allen, Valery A. Petrushin, Damian Roqueiro, and Gang Wei. Semantic web techniques for searching and navigating video shots in BBC rushes. In TRECVID 2006 Workshop, 2006