BIOINFORMATICS TOOLS FOR EXPLORING REGULATORY MECHANISMS

Guosheng Zhang

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum in Bioinformatics and Computational Biology in the School of Medicine.

Chapel Hill
2016

Approved by:

David Neil Hayes

Leslie Lange

Yun Li

Karen Mohlke

Wei Sun

ABSTRACT

Guosheng Zhang: Bioinformatics Tools for Exploring Regulatory Mechanisms
(Under the direction of Yun Li)

Gene expression is the fundamental initial step in the flow of genetic information in biological systems and it is controlled by multiple precisely coordinated regulatory mechanisms, such as structural and epigenetic regulations. Dysregulation of gene expression plays important roles in the development of a broad range of diseases. Modern high-throughput technologies provide unprecedented opportunities to investigate these diverse regulatory mechanisms on a genome-wide scale. Here we develop several methods to analyze these omics profiles.

First, Hi-C experiments generate genome-wide contact frequencies between pairs of loci by sequencing DNA segments ligated from loci in close spatial proximity. To detect biologically meaningful interactions between loci, we propose a hidden Markov random field (HMRF) based Bayesian method to rigorously model interaction probabilities in the two-dimensional space based on the contact frequency matrix. By borrowing information from neighboring loci pairs, our method demonstrates superior reproducibility and statistical power in both simulation studies and real data analysis.

Second, DNA methylation is a key epigenetic mark involved in both normal development and disease progression. To facilitate joint analysis of methylation data from multiple platforms with varying resolution, we propose a penalized functional regression model to impute missing methylation data. By incorporating functional predictors, our model utilizes information from non-local probes to improve imputation quality. We compared the performance of our functional model to linear regression and the best single probe surrogate in real data and via simulations, and our method showed higher imputation accuracy. The simulated association study further demonstrated that our method substantially improves the statistical power to identify trait-associated methylation loci in epigenome-wide association study (EWAS).

Finally, we applied an integrative analysis to characterize molecular systems associated with hepatocellular carcinoma (HCC). Dysregulaton of inflammation-related genes plays a pivotal role in the development of HCC. We performed array-based analyses to comprehensively investigate the contributions of DNA methylation and somatic copy number aberration (SCNA) to the aberrant expression of inflammation-related genes in 30 HCCs and paired non-tumor tissues. The results were validated in public datasets and an additional sample set of 47 paired HCCs and non-tumor tissues. We found that DNA methylation and SCNA together contributed to less than 30% aberrant expression of inflammation-related genes, suggesting that other molecular mechanisms might play major role in the dysregulation in HCCs.

To my family and friend, I couldn’t have done this without you.
Thank you for all of your support along the way.

ACKNOWLEDGEMENTS

First of all I would like to offer countless thanks to my adviser, Dr. Yun Li, for her tireless support and encouragement, and each of my committee members Wei Sun, David Neil Hayes, Karen Mohlke and Leslie Lange for their thoughtful contributions to this research. I would also like to thank my collaborators from UNC-Chapel Hill and other organizations for their invaluable comments and suggestions: Ming Hu from New York University Department of Population Health; Jian Kang from University of Michigan Department of Biostatistics; Fulai Jin for sharing the Hi-C data used in chapter 3; Karen N. Conneely for providing the DNA methylation data used in chapter 4; Dianke Yu and Dongxin Lin for providing the data and helping to validate the results in chapter 5. Finally, I would like to thank my co-workers, friends and family for their patience and support of this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AFC	Anchor-Fragment Caller
AML	Acute Myeloid Leukemia
CCLE	Cancer Cell Line Encyclopedia
CGI	CpG Island
CNV	Copy Number Variation
DNMT	DNA Methyltransferases
ENCODE	Encyclopedia of DNA Elements
ESC	Embryonic Stem Cell
EWAS	Epigenome-wide Association Study
FDR	False Discovery Rate
FPR	False Positive Rate
GEO	Gene Expression Omnibus
GWAS	Genome-wide Association Study
HBV	Hepatitis B Virus
HCC	Hepatocellular Carcinoma
HCV	Hepatitis C Virus
HM27	Illumina HumanMethylation27 BeadChip
HM450	Illumina HumanMethylation450 BeadChip
HMM	Hidden Markov Model
HMRF	Hidden Markov Random Field
LD	Linkage Disequilibrium
LOESS	Locally Weighted Scatterplot Smoothing

LRR	Log R Ratio
MSE	Mean Squared Error
PCR	Polymerase Chain Reaction
PLS	Partial Least Squares
QC	Quality Control
RR	Recovery Rate
SCNA	Somatic Copy Number Aberration
SNP	Single Nucleotide Polymorphism
SVM	Support Vector Machine
TAD	Topologically Associating Domain
TCGA	The Cancer Genome Atlas
TFBS	Transcription Factor Binding Site
TNF	Tumor Necrosis Factor
TSS	Transcription Start Site

CHAPTER 1 MOTIVATION AND BIOLOGICAL JUSTIFICATION

In this document, we will discuss bioinformatics methods for exploring various regulatory mechanisms of gene expression. This section provides an overview of the biological problems we are interested in.

1.1 Gene Expression and Regulation

Gene expression is fundamental in the process by which static DNA information translates into dynamic phenotypes. Understanding gene expression is essential for interpreting functional elements in the genome, identifying molecular systems in cells, and understanding the developmental process. The gene expression programs that establish specific cell states during cell differentiation are controlled by multiple regulatory mechanisms. Dysregulation of these gene expression programs plays important roles in the development of a broad range of diseases.

The transcriptional regulation can be roughly grouped into four main categories. First, structural modifications of DNA, ranging from global change of packing density to local copy number aberrations, have great influence on the transcription. Second, epigenetic modification also plays an important role in transcription. For example, DNA methylation and histone deacetylation usually cooperate and result in gene silencing, while histone acetylation opens up the packed nucleosome, allowing transcription to proceed. Third, the transcription machinery can be modulated by proteins such as transcription factors [Spitz and Furlong 2012] and repressors. Finally, additional control is provided by post-transcriptional regulations, including 5' capping, addition of poly-adenylated tail and microRNA degradation. In this document we will focus on three types of regulations, chromosome structure, somatic copy number aberrations and DNA methylation.

1.2 Chromosome Structure

Chromosomal DNA must be packed nearly three orders of magnitude to fit within the limited space of the nucleus. The intricate, highly compacted folding of the chromosomes, however, is by no means random. Recent technological advances have made it possible to delineate the three-dimensional (3D) organization of the genome [de Wit and de Laat 2012]. It is becoming increasing clear that chromatin architectures are intimately linked to transcription regulation by influencing how genetic information is accessed, read, and interpreted in a given cell and under certain micro-environmental conditions via dynamic interactions among genes and their regulatory elements [Splinter and de Laat 2011].

One prominent feature of the genome is the formation of chromosomal domains at multiple scales [Gibcus and Dekker 2013]. At ~1-10Mb scale, compartments of transcriptionally active euchromatin and inactive heterochromatin are spatially segregated [Lieberman-Aiden, et al. 2009; Simonis, et al. 2006]. Recently, the genome was further shown to be subdivided into conserved topologically associating domains (TADs) at the subcompartment level, which can be hundreds of kilobases in size [Dixon, et al. 2012; Nora, et al. 2012]. A specific locus tends to interact much more frequently with loci located within the same TAD, compared to loci located outside its TAD. The boundaries between TADs are genetically encoded, which was demonstrated by partial fusion of two neighboring TADs in the X-chromosome inactivation center after deletion of the boundary between them [Nora, et al. 2012]. Genome-wide analysis of boundary regions indicated that they are enriched in CTCF binding sites, suggesting that CTCF binding sites may act as a boundary element [Gaszner and Felsenfeld 2006; Phillips and Corces

2009]. Within TAD, a long-range loop structure can be formed to link a distant enhancer with its target gene to regulate gene transcription. Hence, characterization of the 3D genome organizations is critical to understanding how the chromatin structure influences cell fate determination during development [Dekker, et al. 2013; Sajan and Hawkins 2012].

1.3 Somatic Copy Number Aberration

Cancer genesis and progression are enabled through the stepwise accumulation of genomic alterations in somatic cells, including point mutations, somatic copy number aberrations (SCNA), and fusion events, that affect the function of critical genes regulating cell proliferation and apoptosis [Hanahan and Weinberg 2011]. Recurrent genomic abnormalities provide a selection advantage by targeting genes vital for tumorigenesis and metastasis [Albertson, et al. 2003]. Discovery and functional assessment of oncogenes and tumor suppressor genes being targeted by these abnormalities has spurred both the understanding of tumorigenesis and the identification of novel therapeutic targets [Stratton, et al. 2009]. Genes targeted by SCNA, in particular, play central roles in oncogenesis as SCNA results in altered expression of these genes (dosage effect) [Santarius, et al. 2010]. A recent work studied patterns of SCNA across 26 cancer types, and found 24 gains and 18 losses on average per tumor sample [Albertson, et al. 2003]. A literature review reveals numerous examples of genes identified as targets of focal or chromosomal arm-level amplification or deletions. Most notable examples include amplified oncogenes *ERBB2* [Slamon, et al. 1989], *MYC* [Alitalo, et al. 1983], *CAD* [Schimke, et al. 1978; Wahl, et al. 1979], *BCR-ABL* [Koivisto, et al. 1997] and *AR* [Gorre, et al. 2001], and deleted tumor suppressor genes *PTEN* [Li, et al. 1997], *CDKN2A* [Orlow, et al. 1995], *RB1*, *BRCA1*, *BRCA2*, *PTPRJ* and *TP53* [Cavenee, et al. 1983; Nagai, et al. 1994].

1.4 DNA Methylation

Epigenetic modifications are non-sequence changes in chromatin that inherited through cell division, which are often dynamic and tissue-specific [Cedar and Bergman 2012; Scarano, et al. 2005]. DNA methylation is an important epigenetic marker involved not only in normal development [Smith and Meissner 2013] but also in risk and progression to many diseases [Bergman and Cedar 2013]. DNA methylation is a process by which a methyl group is added to the cytosine residue. DNA methylation typically occurs in the context of CpG dinucleotide in the genome and it is mediated by DNA methyltransferases (DNMTs). It has been shown to play a key role in the regulation of gene transcription, X-inactivation, cellular differentiation, and other critical processes such as aging [Bird 2002; Gonzalo 2010].

Due to an increased vulnerability of methylated cytosines to mutate, CpG sites occur with a much lower frequency in the genome than would be expected [Tost 2010]. 70~80% of CpG sites are highly methylated across the genome, which might facilitate chromatin arrangements to repress the transcription of repeated regions, such as Alu sequences and transposons [Bestor and Tycko 1996; Jones 2012]. The methylated regions are also shown to inhibit CTCF binding, allowing the downstream enhancers to activate gene expression [Valenzuela and Kamakaka 2006]. CpG islands (CGIs) are regions with a GC content greater than 50% [Law and Jacobsen 2010; Tost 2010]. CGIs account for 1~2% of the genome and are typically located in promoters of genes, particularly housekeeping genes, in the genome [Shen, et al. 2007]. Unlike other CpG sites, CGIs are mostly unmethylated if the genes are expressed. For example, ~84% of CpG sites in CGIs were unmethylated in human brain [Maunakea, et al. 2010]. CGIs are found to be enriched for transcription factor binding motifs, and the binding of these transcription factors protects CGIs from de novo methylation [Brandeis, et al. 1994; Choy, et al. 2010; Deaton and Bird 2011; Dickson, et al. 2010; Macleod, et al. 1994; Teschendorff, et al. 2009]. Across the genome, DNA methylation levels have been shown to be correlated with chromatin modifications such as histone methylation [Hawkins, et al. 2010; Meissner, et al. 2008; Weber, et al. 2007], cis-regulatory elements, and proximal sequence pattern [Das, et al. 2006; Shen, et al. 2007], indicating the interaction between DNA methylation and cellular phenotypes.

An increasing number of tumor genes are being identified as hypermethylated in normally unmethylated promoter CGIs [Baylin, et al. 1998; Jones and Laird 1999]. This hypermethylation, as well as mutations, is critical for loss-of-gene function and tumorigenesis. Almost half of the tumor suppressor genes can be inactivated by promoter hypermethylation [Baylin, et al. 1998]. Another relevant study shows that most methylation alterations in colon cancer occur not in promoter CGIs, but in sequences up to 2 kb distant, or ‘CpG island shores’ [Irizarry, et al. 2009].

CHAPTER 2 LITERATURE REVIEW

This section presents a partial review of many of the papers previously published on the topic of methods to study regulatory mechanisms. It is by no means complete since the number of these papers is quite large; however, it is an attempt to show the development of methods used to tackle these problems.

2.1 Early Methods for 3C-based Data Analysis

Recent advancements in chromosome conformation capture (3C) [Dekker, et al. 2002] and derived methods (such as 3C, 4C, 5C and Hi-C) allow the study of 3D chromosome organization with increasing resolution and throughput. These 3C-based methods quantify the interaction or contact frequency, how often any pair of loci in the genome is in close spatial proximity. For 3C, a locus is the unit of analysis and corresponds to one restriction enzyme fragment (hereafter termed fragment). Approaches to analyze interaction frequencies fall largely into two complementary categories: *3D model reconstruction* and *peak calling*. The first set of methods simultaneously model contact frequencies of all pairs of loci in the genome to reconstruct 3D structure [Bau, et al. 2011; Hu, et al. 2013; Jhunjhunwala, et al. 2008; Marti-Renom and Mirny 2011; Russel, et al. 2012; Trieu and Cheng 2014]. The second set of methods aim to identify interaction peaks, meaning pairs of loci where the observed contact frequency is higher than expected from non-random chromatin looping or co-location events [Ay, et al. 2014; Duan, et al. 2010; Sanyal, et al. 2012]. To answer many important biological questions (e.g., pinpointing individual *cis*-regulatory elements), higher resolution for the contributing loci is highly desirable, if not indispensable.

This section focuses on peak calling. Several computational and statistical methods have been developed for this important peak calling task for data generated from 3C-based methods. Sanyal et al. [Sanyal, et al. 2012] developed a 5C peak calling algorithm where they first estimated the null contact frequencies (average and standard deviation) using nonparametric lowess smoothing over genomic distance (using all pairs with the assumption that the vast majority of interactions are random collisions), then calculated standardized z-scores and raw p-values by fitting the z-scores to a Weibull distribution, followed finally by converting the raw p-values into q-values for FDR analysis. Duan et al. [Duan, et al. 2010] binned pairs of loci according to genomic distance, estimated null contact probabilities within each bin, and called peaks by assuming the contact frequency of every pair in each bin followed an identical binomial distribution. Jin et al. [Jin, et al. 2013] developed a pipeline to estimate the expected contact frequency accounting for locus length, inter-locus distance, mappability, and GC content, and then tested for significant interaction by assuming the observed contact frequency followed a negative binomial distribution. Most recently, Ay et al. [Ay, et al. 2014] refined the binning method in Duan et al. [Duan, et al. 2010] and to develop Fit-Hi-C. Specifically, Fit-Hi-C provided more accurate estimates of the contact probabilities by fitting nonparametric spline curves across genomic distances (instead of discrete binning), re-fitting spline curves after filtering non-random collisions based on the initial spline, and modeling other Hi-C biases by incorporating locus-specific correction factors inferred from a previously published iterative correction and eigenvector decomposition method [Imakaev, et al. 2012].

These existing methods have advanced the field by improving the accuracy in the estimation of the expected contact frequencies under the null (i.e., random collisions). All these methods take into account genomic distance between the pair of loci under inference during the estimation, with some [Ay, et al. 2014; Jin, et al. 2013] incorporating other genomic biases. However, all existing methods, by testing each individual pair of loci independently, ignore the potential correlation among pairs of loci. This was less of an issue with lower resolution data when multiple fragments combined into meta-fragments served as the units of analysis.

When analyzing a fragment resolution Hi-C data, Jin et al. [Jin, et al. 2013] recognized this potential issue and developed the anchor-fragment caller (AFC), an *ad hoc* approach to accommodate the correlation of peak status among neighboring fragment pairs. In AFC, one anchor was fixed (either a fragment or mega-fragment from consecutive smaller fragments) and one-dimensional peak calling was performed. For each anchor, the algorithm started with the identification of candidate peak regions. A candidate peak region could encompass multiple consecutive fragments with moderate marginal evidence for non-random interaction with the anchor and, importantly, AFC allows for small gaps. Peaks were called by aggregating information across the entire candidate peak region via assigning thresholds on read counts and p-values from contributing fragment pairs as well as from the entire region cumulatively. As an initial attempt to model the spatial dependency of the underlying peak status, AFC performed reasonably.

2.2 Early Methods for DNA Methylation Imputation

Imputation has been successfully employed in many genetic, genomic and epigenomic contexts [Donner, et al. 2012; Ernst and Kellis 2015; Jewett, et al. 2012; Li, et al. 2009; Zhang, et al. 2015]. Among them, DNA methylation status is difficult to predict since it is tissue-specific and poorly conserved. Multiple methods have been proposed to impute methylation levels across tissue types [Ma, et al. 2014] or employing various genomic and epigenomic features [Bock, et al. 2006; Das, et al. 2006; Zhang, et al. 2015].

Most existing methods formulate the methylation prediction as a binary classification problem, i.e., distinguish between methylated and unmethylated CpG sites [Fan, et al. 2008; Fang, et al. 2006]. Related methods conduct methylation imputation at two levels – CGI (or CpG-rich region) or CpG dinucleotide. At the CGI level, average methylation status for windows of the genome is predicted, and some studies achieve a > 90% accuracy. However, as we mentioned earlier, most CpG sites residing within CGIs tend to remain unmethylated across the whole genome [Jones 2012]. Thus it is not surprising that these methods can achieve a high accuracy in these regions. At the CpG dinucleotide level, methylation status is predicted for individual CpG sites, and usually the accuracy is lower [Bhasin, et al. 2005]. A few studies predict methylation levels as a continuous variable, but instead of genome-wide analysis, the prediction is limited to specific genomic regions.

A key step for building these predictive models is to select features. The features can be roughly grouped into two categories: genetic and epigenetic. Genetic features include (1) proximal DNA sequence patterns, (2) distribution patterns of functional and evolutionarily conserved elements [Siepel, et al. 2005], such as transcription factor binding sites (TFBSs), (3) DNA structure (e.g., co-localized introns), (4) GC contents, and (5) functional annotation of nearby genes. Epigenetic features mainly include methylation and acetylation status of the histones. Several studies used only DNA composition [Bhasin, et al. 2005; Das, et al. 2006; Zhou, et al. 2012], or methylation levels of the same CpG sites from different tissues as features [Ma, et al. 2014], while some used several hundred features, including DNA composition, DNA structure, repeat elements, and number of SNPs [Bock, et al. 2006; Zheng, et al. 2013]. However, it is difficult to quantify the importance of each feature across these studies due to the different methods and prediction objectives.

The majority of these methods are based on support vector machine (SVM) classifiers using linear kernel [Bhasin, et al. 2005; Bock, et al. 2006; Das, et al. 2006; Fan, et al. 2008; Fang, et al. 2006; Ma, et al. 2014; Previti, et al. 2009; Zheng, et al. 2013; Zhou, et al. 2012], where non-additive interactions between features are not modeled. Several studies found that decision tree [Previti, et al. 2009], random forest [Zhang, et al. 2015], or naïve Bayes classifier can achieve better prediction performance.

2.3 Early Methods for Integrative Analysis

Biological system is subjected to precisely coordinated controls at multiple layers, across the levels of genetic, epigenetic, transcriptional, and translational regulations. Different layers intertwined to form multiple complex and extensively coupled networks [Maniatis and Reed 2002; Moore 2005; Orphanides and Reinberg 2002]. Recent development of various high-throughput technologies has enabled researchers to collect diverse information on the same set of samples. Microarray and next-generation sequencing technologies are used to profile genome-wide gene expression levels, genetic variations (e.g. SNP and CNV), epigenetic regulation (e.g. DNA methylation and histone modifications), post-transcriptional regulation (e.g. microRNA expression), and protein abundance. The amount of available biological data from multiple technologies and platforms is expanding rapidly. Large online databases such as the UCSC Genome Browser [Speir, et al. 2016] and large-scale projects such as The Cancer Genome Atlas (TCGA) [Cancer Genome Atlas Research 2008] often contain multiple data types collected from a cohort of samples.

For different data types measured on the same set of samples, methods to analyze them separately are well established. However, individual analyses will miss the critical associations between data types. Multi-omics datasets provide an opportunity to discover the interplay across these different regulatory layers. Since different data types have different scales and units, we cannot simply combine them for analysis, thus novel computational methods are needed to explore associations among multiple data types and aggregate these multiple data sources when making inference about the samples.

Earlier relevant efforts focused on analyzing two-dimensional genomic datasets. For example, eQTL method can jointly analyze SNP and gene expression data to identify target genes of regulatory SNPs [Zhang, et al. 2010]. General methods such as partial least squares (PLS) were also used to examine the relation between two sets of variables. However, the restriction of these methods to pairwise comparisons limits its utility in examining relations among more than two data types. Recently, several methods have been developed to analyze genomic datasets with more than two data types. For example, multiple canonical correlation analysis (mCCA) was introduced as an extension of PLS to more than two data types [Witten and Tibshirani 2009], and it can explore associations and structures on multi-omics data in a supervised manner. Several unsupervised methods were also proposed. For example, iCluster incorporates a joint latent variable model for integrative clustering and tumor subtype discovery [Shen, et al. 2009]. An adaptive clustering approach was also developed to integratively analyze genome-wide gene expression, DNA methylation, microRNA expression, and copy number alteration profiles [Zhang, et al. 2013a].

Integrative analysis is widely used in cancer studies, due to its importance and molecular complexity. Genomic data at high resolution have been collected for thousands of tumors from large-scale projects such as TCGA and Cancer Cell Line Encyclopedia (CCLE) [Barretina, et al. 2012]. Systematic characterization of cancer genomes has revealed considerable heterogeneity. Each tumor is unique and different individuals typically harbor diverse genomic aberrations occurring in different genes, of which only a few drive cancer proliferation and metastasis. Thus, identifying those driver genes with functional importance and therapeutic implications, and distinguishing them from passenger genes has emerged as a major challenge in the genomic characterization of cancer. Multiple studies demonstrate that the considerable heterogeneity can be resolved by an integrative analysis [Akavia, et al. 2010; Shen, et al. 2009]. Another finding is that genomic alterations among individuals may differ, but usually common pathways are involved. Based on this finding, PARADIGM identifies relevant pathways involved in tumorigenesis from patient-specific genomic alterations [Vaske, et al. 2010]. This method modeled each gene as a factor graph with a set of interconnected nodes encoding copy number, expression, protein activity and so on, allowing the incorporation of many data types.

CHAPTER 3 A BAYESIAN MODEL FOR THE DETECTION OF LONG-RANGE CHROMOSOMAL INTERACTIONS IN HI-C DATA

3.1 Introduction

Identifying non-random contacts in the 3D genome organization is of fundamental biological interest to researchers due to their relevance for functional regulation. For instance, it can shed light on the functional mechanisms of non-coding complex trait associations identified in genome-wide association studies (GWAS). GWAS have been resoundingly successful, identifying thousands of variants associated with complex traits. However, only a small proportion (7~12%) fall in protein coding regions [Hindorff, et al. 2009; Kumar, et al. 2012; Pennisi 2011; Ward and Kellis 2012], making interpretation of non-coding variants imperative. Although a large number of regulatory elements have been annotated [Consortium 2012; Maurano, et al. 2012], their target genes are largely unknown [Jin, et al. 2013; Niu, et al. 2014]. Recent 3C-based studies are generating an increasingly comprehensive catalog of interactions between genes and their regulatory elements in different cell types at varying resolution across multiple organisms including drosophila, yeast, mouse, and human [Hou, et al. 2012; Lieberman-Aiden, et al. 2009; Sexton, et al. 2012; Smallwood and Ren 2013]. Such information will be fundamental to understanding functional mechanisms. For example, a recent study [Smemo, et al. 2014] used 4C data to identify long-range (at megabase distances) interactions between the obesity-associated intronic variants in *FTO* and the homeobox gene *IRX3*, with the expression of *IRX3* rather than *FTO* being directly linked to body mass. This study showcased the value of interactions identified from the 3C-based studies for shedding light on the functional mechanisms of genetic variants implicated by GWAS.

Several computational and statistical methods have been developed for peak calling for data generated from 3C-based methods. However, we believe that the existing methods are not yet optimal, and that improvements in multiple aspects are needed. First, an improved analysis suite for data from 3C-derived methods should be based on an explicit model that yields clear and reproducible expectations for genome-wide interaction frequencies. Second, existing approaches choose anchor fragment(s) arbitrarily and also ignore any correlations between neighboring fragments or anchors. For example, we found that neighboring anchors often interact with the same target fragments, suggesting that these anchors are parts of a bigger region involved in the same DNA looping event. Therefore an ideal peak caller should consider correlations between neighboring fragments in the context of a two-dimensional (2D) contact matrix generated from 3C-derived technologies. Third, one-dimensional calling approaches are not optimal, do not incorporate useful existing information, and considerable benefits can be gained using a 2D approach. For example, we observed AFC asymmetric peak calls and lower power in the identification of non-random interactions (details in Results section). Thus, these observations motivated us to develop rigorous statistical models that efficiently use information from neighbors in the 2D space.

Here we present a hidden Markov random field (HMRF) based Bayesian model for peak calling using Hi-C data. Our approach improves on previous methods by explicitly borrowing information from neighboring fragment pairs via modeling the dependency in the 2D space. Our results in real data and from extensive simulations indicate superior performance of our method over existing methods, across a range of underlying dependency structure.

3.2 Methods

3.2.1 Notations

Hi-C generates a contact frequency matrix between pairs of fragments. Assume a total of N fragments under consideration. Let u_{ij}, $1 \leq i < j \leq N$ denote the observed contact frequency between fragment i and fragment j. Similarly, let e_{ij}, $1 \leq i < j \leq N$ denote the expected contact frequency between fragment i and fragment j under random collisions. Let the binary indicator variable Z_{ij} take two possible values 1 and -1 which represent the peak status underlying fragment pair i and j, with $Z_{ij} = 1$ corresponding to a peak (i.e., a non-random interaction) and $Z_{ij} = -1$ corresponding a non-peak (i.e., a random collision event).

3.2.2 Mixture of Negative Binomials

We assume that the observed contact frequencies u_{ij} follow a negative binomial distribution, $u_{ij} \sim NB(\mu_{ij}, \phi)$, where ϕ is the over-dispersion parameter and u_{ij} has mean μ_{ij} and variance $\mu_{ij} + \mu_{ij}^2/\phi$. The benefit of using a negative binomial distribution (over Poisson or binomial distribution) is its allowance for over-dispersion, often observed in Hi-C data [Jin, et al. 2013].

Furthermore, we assume that the observed contact frequencies follow a mixture of negative binomial distributions as a consequence of the mixture of underlying interaction status Z_{ij}'s. Specifically, let $\theta > 0$ represent the peak to background ratio (signal to noise ratio). We assume the following on $log\, \mu_{ij}$:

$$log\, \mu_{ij} = \begin{cases} log\, e_{ij} + \theta, & Z_{ij} = 1 \\ log\, e_{ij}\,, & Z_{ij} = -1 \end{cases}$$

where e_{ij}'s are expected counts under random collision events, estimated using existing methods such as ICE [Imakaev, et al. 2012] or Fit-Hi-C [Ay, et al. 2014]. Thus we use the following negative binomial mixture distribution:

$$u_{ij} \sim NB\left(e_{ij}e^{\theta(Z_{ij}+1)/2}, \phi\right).$$

3.2.3 Hidden Markov Random Field (HMRF) Model

A HMRF is a generalized hidden Markov model (HMM) in a higher dimensional space [Besag, et al. 1995]. Instead of an underlying Markov chain in HMM, HMRF has an underlying Markov random field, a set of random variables having a Markov property described by an undirected graph. HMRF has been applied in genetics, including evaluation of population structure [François, et al. 2006], gene expression data [Stingo and Vannucci 2011], network-based genomic discovery [Wei and Pan 2010], and GWAS [Li, et al. 2010]. We use HMRF to account for the local spatial dependency among adjacent fragment pairs, and simultaneously detect all 2D peaks by borrowing information from neighboring fragment pairs. Our HMRF modeling is conceptually similar to the employment of HMM or Bayesian hidden Ising model for peak identification from ChIP-Seq data [Choi, et al. 2010; Mo 2012; Qin, et al. 2010], but we extend the modeling from a one-dimensional space to a two-dimensional space. In our HMRF model, we adopt the following Ising prior [Kindermann, et al. 1980] for the binary variable $Z_{ij} \in \{-1,1\}$ representing the unobserved peak status underlying fragment i and fragment j such that Z_{ij} only depends on the status of four neighboring fragment pairs $(i+1,j)$, $(i-1,j)$, $(i,j+1)$ and $(i,j-1)$:

$$\pi\left(Z_{ij}|\psi\right) = \frac{\exp\left\{\psi Z_{ij} \sum_{|i'-i|+|j'-j|=1} Z_{i'j'}\right\}}{W(\psi)}, \qquad 1 \le i < j \le N,$$

where ψ is the inverse temperature parameter measuring the level of clustering among Z_{ij}'s. The term $W(\psi)$ is the normalizing function ensuring the probability mass sum to 1. The case $\psi = 0$ corresponds to independent uniform prior on Z_{ij}'s, analogous to the disordered states at infinite temperature. Large values of ψ correspond to more tightly clustered configurations of Z_{ij}'s, analogous to more ordered/correlated states at low temperature. In Hi-C data, positive clustering is expected, particularly with the high-resolution Hi-C data where neighboring fragment pairs are likely to share the underlying peak or non-peak status. Our model explicitly models the level of clustering and estimates the value of ψ based on data [Besag, et al. 1995]. Although our model is expected to manifest its advantages more with clustered hidden states, but, even in the special case of no clustering, it is unlikely that our model will incur any power loss if the inverse temperature parameters can be calibrated to its true value (close to 0 in this case). Figure 3.1 shows the histogram of domain-specific inverse temperature estimates from real data at fragment pair level (based on the peaks reported by AFC), which clearly suggest non-negligible clustering of the peaks for pairs within topological domains [Dixon, et al. 2012; Hou, et al. 2012; Li, et al. 2012; Nora, et al. 2012]. In this work, we focus on the detection of intra-domain interactions, which account for the vast majority of non-random interactions (e.g., 95.3% interactions reported by Jin et al. [Jin, et al. 2013] are intra-domain). We followed domain definitions from Dixon *et al.* 2012 [Dixon, et al. 2012].

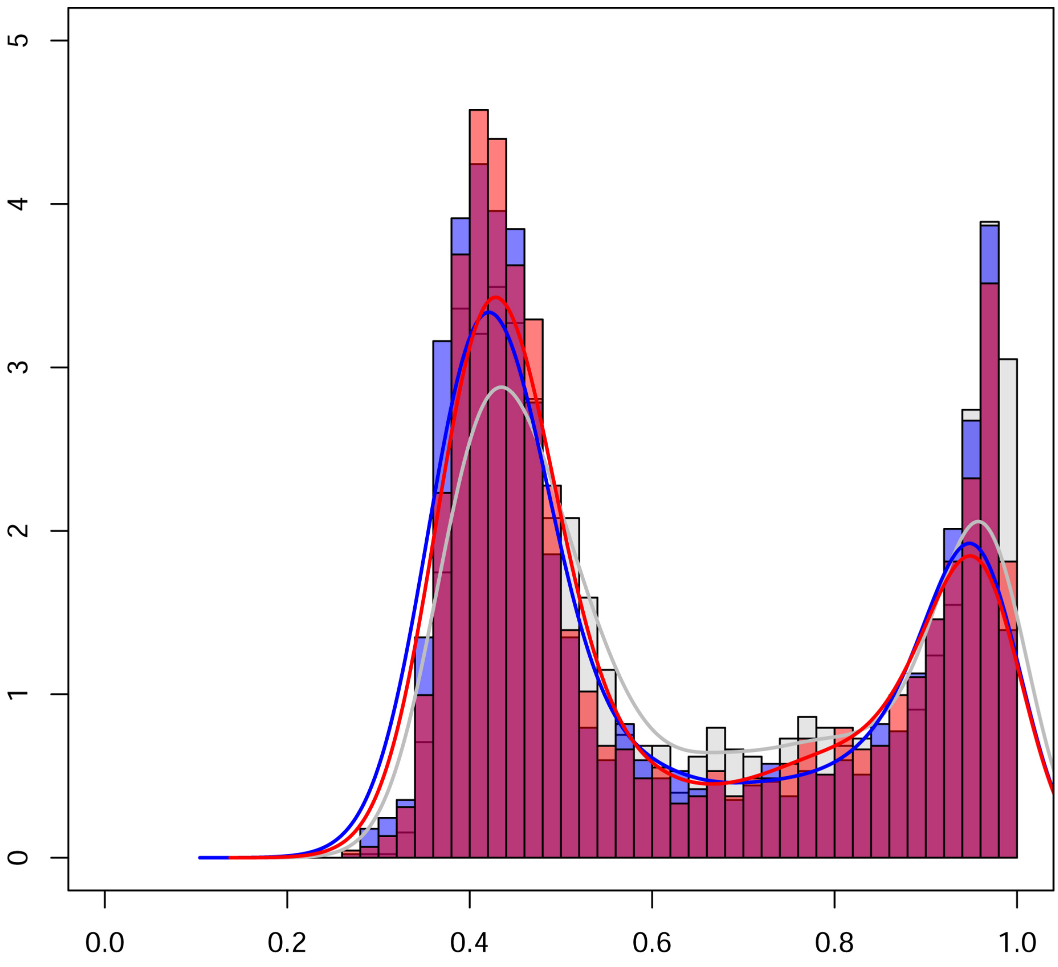


Figure 3.1 Overlapping histograms of inverse temperature ψ estimates in combined (gray), before (blue) and after (red) TNF-α treatment.

3.2.4 Bayesian Inference and the Joint Probability

We adopt a Bayesian approach [Gelman 2004] for parameter inference where the inference is based on the posterior distributions. We will start with specifying the priors. For convenience and computational efficiency, we make a re-parameterization: $\gamma = \phi^{-1}$. By default, our model uses weak priors with large variance: a translated gamma distribution for θ : $\pi(\theta) = Gamma(\theta - \theta_0; 2, 2)$, a gamma distribution for γ: $\gamma{\sim}Gamma(\gamma; 0.1,1)$ and a uniform distribution for ψ: $\psi{\sim}Unif(0,1)$. To evaluate the impact of priors, we considered other priors and found little impact on final peaks called (Spearman correlation > 0.99 on average). Note that

θ_0 is fixed to ensure model identifiability (we use $\theta_0 = 0.5$ by default). The likelihood is fully specified based on the mixture of negative binomial distributions introduced earlier. Combined with the conditional independence assumption of u_{ij} given Z_{ij}, we have

$$\pi(\{u_{ij}\}|\{Z_{ij}\},\theta,\gamma) = \prod_{1\le i<j\le N} \left(\frac{\frac{1}{\gamma}}{\frac{1}{\gamma} + e_{ij}e^{\theta(Z_{ij}+1)/2}}\right)^{\frac{1}{\gamma}} \frac{\Gamma\left(\frac{1}{\gamma} + u_{ij}\right)}{\Gamma\left(\frac{1}{\gamma}\right) u_{ij}!} \left(\frac{e_{ij}e^{\theta(Z_{ij}+1)/2}}{\frac{1}{\gamma} + e_{ij}e^{\theta(Z_{ij}+1)/2}}\right)^{u_{ij}}.$$

The posterior probability can be written as

$$\pi(\{Z_{ij}\},\theta,\gamma,\psi|\{u_{ij}\}) \propto \pi(\{u_{ij}\},\{Z_{ij}\},\theta,\gamma,\psi) = \pi(\{u_{ij}\}|\{Z_{ij}\},\theta,\gamma)\pi(\{Z_{ij}\}|\psi)\pi(\theta)\pi(\gamma)\pi(\psi).$$

We used Metropolis-Hastings algorithm to infer all parameters except the inverse temperature parameter ψ, which is estimated using a pseudo-likelihood approach.

3.3 Results

Comprehensive simulation studies have demonstrated the superior performance of our HMRF Bayesian caller over other available methods. In particular, our simulations showed that our model is able to accurately estimate the inverse temperature parameter ψ across a wide range of spatially dependent patterns and as a result improved power for calling. Next, we showcase the improved reproducibility and statistical power of our method in real data analysis. As aforementioned, our method was motivated by our observations in real data and was developed for re-analysis of the fragment resolution Hi-C data generated by Jin et al. [Jin, et al. 2013]. In the original study, twelve replicates of primary IMR90 human fibroblast cells (including six replicates untreated cells and six replicates after TNF-α treatment) were used to generate ~3.4 billion paired-end reads. This unprecedented sequencing depth allowed direct identification of interacting fragments. One major finding of this study is that TNF-α responsive enhancers are already in contact with their target promoters before signaling, manifested by similar peak patterns observed under each condition separately. Motivated by this finding and the insufficient sequencing depth in each condition (i.e., before or after TNF-α treatment), we combined data from the two conditions to achieve higher statistical power in detecting fragment resolution chromatin interaction.

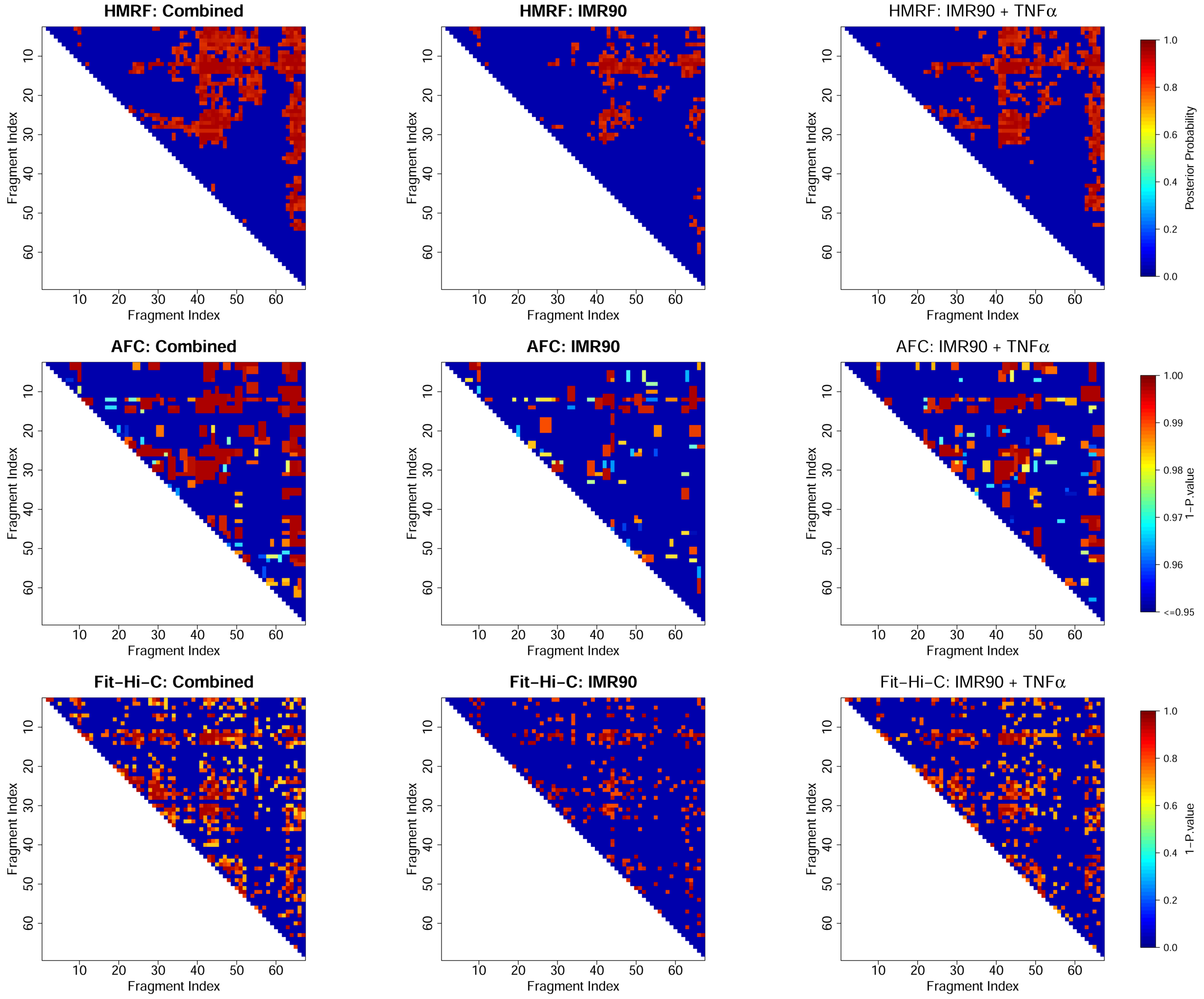


Figure 3.2 Peaks called in the domain Chr.17:29.52Mb-29.72Mb. For each dataset (combined, before, and after TNF-$\boldsymbol{\alpha}$ treatment), the same number of peaks as using AFC method was shown (based on posterior probabilities for HMRF and p-values for Fit-Hi-C) for comparison.

We thus first test our HMRF Bayesian caller in this Hi-C data set. We analyzed three datasets (1) IMR90 before TNF-α treatment, (2) IMR90 after TNF-α treatment and (3) the combined dataset (dataset by pooling data from datasets 1 and 2). Under the rationale that peak patterns of the two conditions are shared, a robust caller is expected to identify similar patterns for the three datasets. Figure 3.2 shows peak calling results from one domain chr17: 29.52Mb-29.72Mb. We observed that fewer peaks were called in datasets 1 and 2, particularly dataset 1 where the total number of reads was 75.1% of that in the dataset 2. Comparatively, our method encourages more clustering of peaks and more consistent results across the three datasets. For example, within this particular domain, 40.6% and 69.9% of the peaks called in the combined dataset were detected using only dataset 1 and 2, respectively, by our method, compared with 35.3% and 65.2% (40.3% and 67.9%) by AFC (Fit-Hi-C). Genome-wide quantitative comparisons are presented below (Table 3.1 and Table 3.2).

Table 3.1 Genome-wide real data evaluation based on False Positive Rate (FPR), False Discovery Rate (FDR) and Recover Rate (RR). Assuming calling result for the combined dataset is the true peak pattern, we summarized the following measures for 1,432 domains, i.e. genome-wide. We reported the genome-wide average of false positive rate (FPR), false discovery rate (FDR) and recovery rate (RR) by the HMRF-Bayesian method, AFC method and Fit-Hi-C for both IMR90 before TNF-$\boldsymbol{\alpha}$ treatment and IMR90 after TNF-$\boldsymbol{\alpha}$ treatment. We found that the HMRF-Bayesian method has better performance than AFC method and Fit-Hi-C.

Dataset	Method	FPR	FDR	RR
IMR90	HMRF-Bayesian	0.52%	15.60%	42.30%
IMR90	AFC	0.60%	18.40%	41.10%
IMR90	Fit-Hi-C	0.64%	19.20%	41.20%
IMR90+TNF-α	HMRF-Bayesian	0.83%	18.50%	55.40%
IMR90+TNF-α	AFC	0.98%	22.40%	52.90%
IMR90+TNF-α	Fit-Hi-C	1.00%	22.50%	53.30%

Dataset	Method	FPR	FDR	RR
Split1	HMRF	0.84%	19.10%	55.90%
Split1	AFC	0.97%	22.60%	54.00%
Split1	Fit-Hi-C	1.03%	23.50%	53.70%
Split2	HMRF	0.47%	15.20%	41.30%
Split2	AFC	0.56%	18.60%	39.90%
Split2	Fit-Hi-C	0.58%	18.80%	40.30%

Table 3.2 Genome-wide real data evaluation based on the consistency measure (Jaccard Index).

Method	IMR90 vs. IMR90+TNF-α*	Split1 vs Split2*
HMRF	22.1±0.33%	22.7%±0.32%
AFC	18.4±0.32%	18.5%±0.31%
Fit-Hi-C	13.7±0.29%	13.6%±0.28%

*Mean±SE

We next proceeded to quantitatively and systematically evaluate the performance based on genome-wide calling for all domains. For a fair comparison, we selected thresholds based on posterior peak probabilities for our method and p-values for Fit-Hi-C to match the number of peaks called by AFC for each dataset. We also performed other comparisons where we matched the number of peaks called by either our method or Fit-Hi-C, or where we let each method call peaks according to its own criterion and found similar patterns. Treating peaks called in the combined dataset as truth, we gauged performance in single-condition datasets using the following three statistics: false positive rate (FPR), false discovery rate (FDR) and recovery rate (RR). Denote the number of false positives, true positives, false negatives and true negatives as

FP, TP, FN and TN, where the truth is defined according to AFC results from the combined dataset and the four numbers sum up to the total number of intra-domain fragment pairs. We have FPR=FP/(FP+TN), FDR=FP/(FP+TP), and RR=TP/(TP+TN). As shown in Table 3.1 upper panel, methods accounting for potential dependency of underlying peak statuses (AFC and our HMRF Bayesian caller) resulted in better performance than Fit-Hi-C which models fragment pairs independently. Furthermore, our method outperformed the others for all three measures. For example, for IMR90 before TNF-α treatment, we obtained FPR=0.52%, FDR=15.6% and RR=42.3% for our HMRF Bayesian caller, compared with FPR=0.60%, FDR=18.4% and RR=41.1% for AFC, and FPR=0.64%, FDR=19.2% and RR=41.2% for Fit-Hi-C. By borrowing information from neighboring fragment pairs in a probabilistic framework, our method lead to more robust inference with simultaneously lower false positive, false discovery rates and higher recovery rate.

In addition, for each caller, we calculated the Jaccard Index [Hamers, et al. 1989] between the peak sets from the two conditions, defined as the ratio of number of peaks identified under both conditions over the number of peaks identified by either. Average Jaccard Index across all domains genome-wide is shown in Table 3.2 for each method. Again, methods accounting for the dependency of underlying peak status show higher concordance across conditions. Average Jaccard Index improved by 33.6% and 61.3% respectively, from 13.7% (Fit-Hi-C) to 18.4% (AFC) and 22.1% (HMRF).

To avoid potential systematic differences between treated and untreated conditions in terms of peak status (although not supported by results in Jin et al. [Jin, et al. 2013]), we also analyzed two randomly split datasets as described by Jin et al. [Jin, et al. 2013]. Results shown in

Table 3.1 (lower panel) and Table 3.2 (rightmost column) similarly show better reproducibility and robustness of our methods over existing ones.

Given one important utility of called peaks is to illuminate biologically meaningful interactions, we directly evaluated the power to identify one important category of biological interactions: between enhancers and transcription start sites (TSS). We used the enhancer-promoter connection map based on multi-tissue correlations between distal and promoter chromatin accessibility [Thurman, et al. 2012], augmented with results from multi-tissue correlations between chromatin accessibility and gene expression [Sheffield, et al. 2013], retrieved from http://dnase.med.unc.edu/supplement/allGeneCorrelations100000.p2.txt.gz. We left Fit-Hi-C out in the comparison as it showed incomparable reproducibility with the other callers. Figure 3.3 demonstrates the increased power of HMRF over AFC with up to 11.3% more enhancer-TSS interactions identified by HMRF, given the same number of peak regions called by two methods. In addition, Figure 3.4 shows one particular example where the potential target gene *CTSB* [Maurano, et al. 2012] of a GWAS variant rs1600249 [Freudenberg, et al. 2011] was missed by AFC but captured by our method.

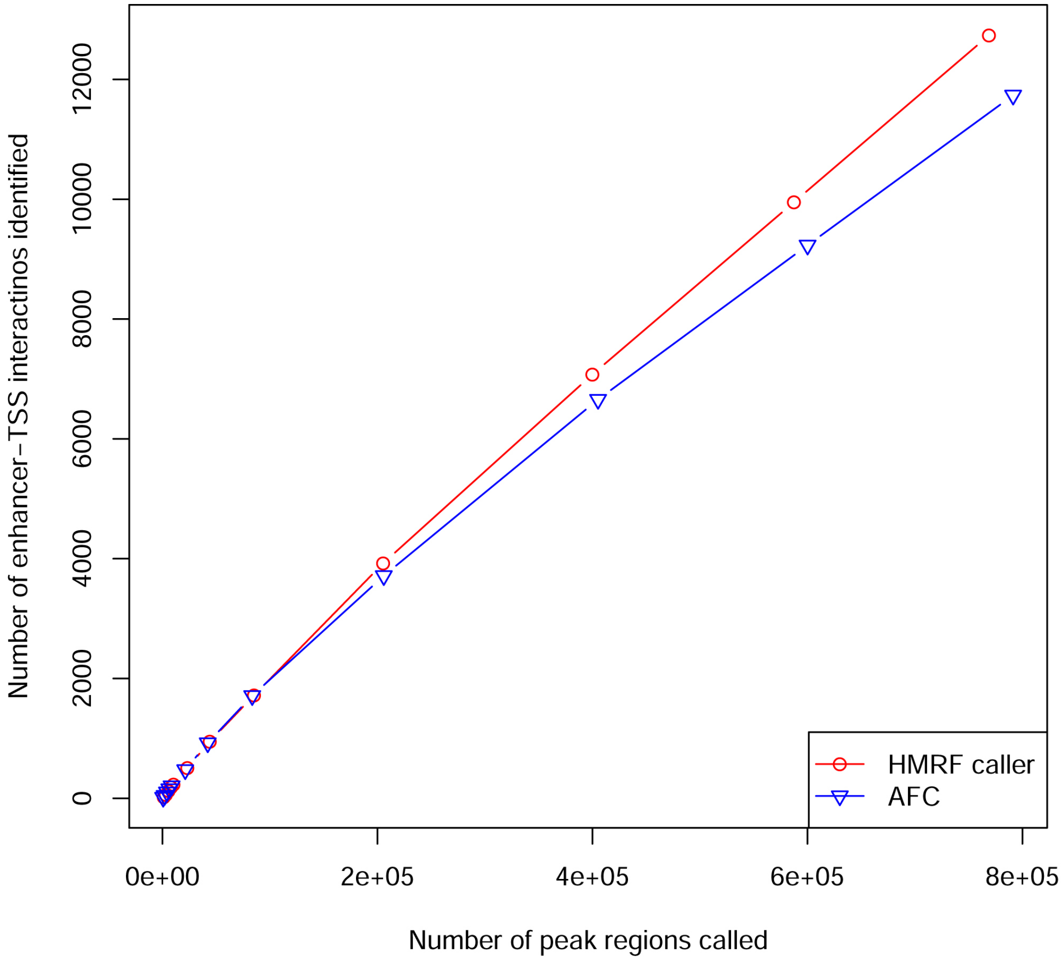


Figure 3.3 Power to identify enhancer-TSS interactions.

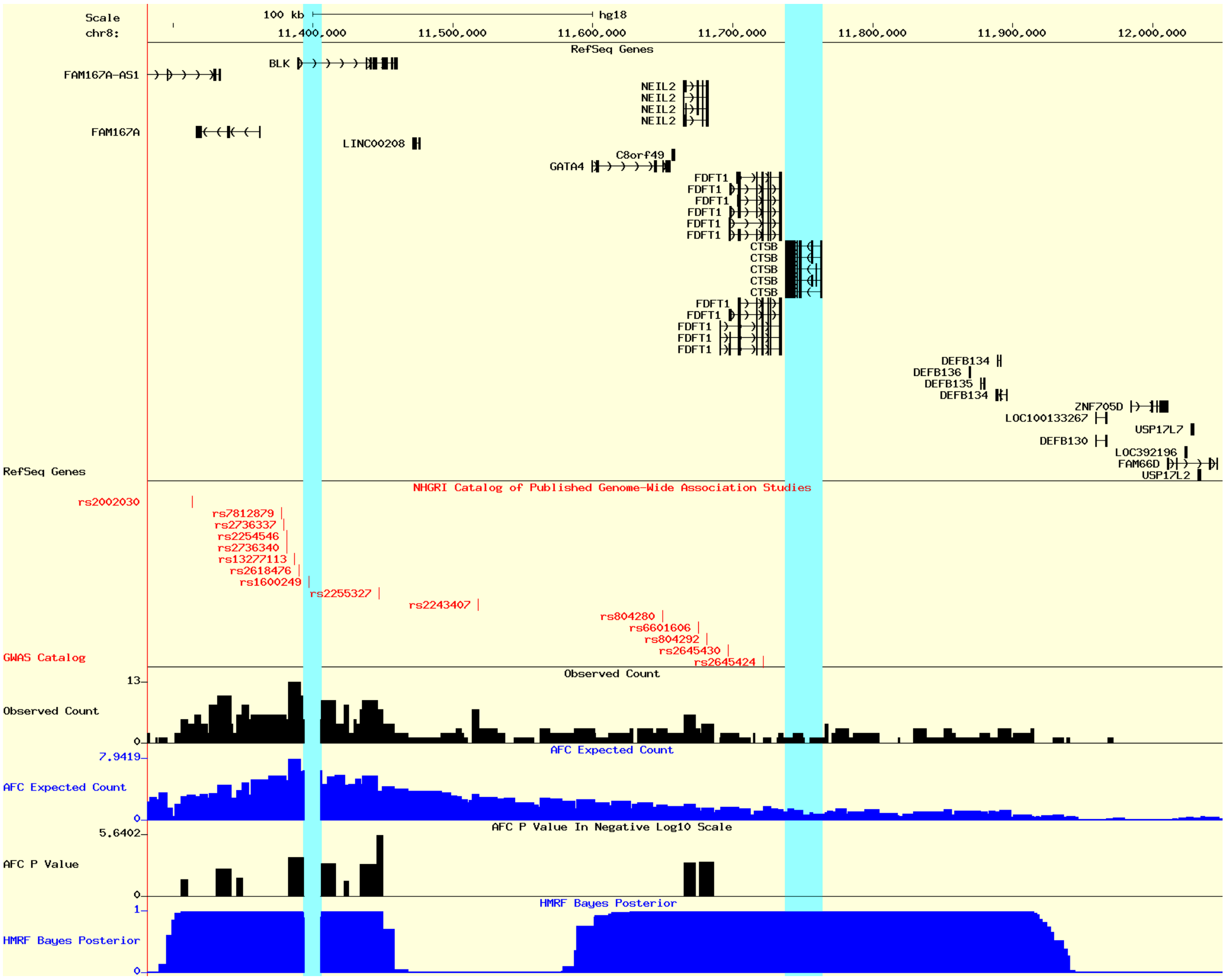


Figure 3.4 *CTSB*, potential target gene for GWAS variant rs1600249.

In addition, we performed transcription factor binding sites (TFBS) and active TSS (reported by Jin, et *al.*, 2013) enrichment analysis to elucidate the biological relevance of identified interactions. Specifically, we evaluated two aspects. First, we tested if fragment pairs detected as interacting loci are enriched with TFBS. Second, we compared the number of interacting loci for TFBS versus non-TFBS. We used ENCODE IMR90 TFBS information retrieved from

http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeRegTfbsClustered

where TFBS were called from ChIP-seq data using the computational pipeline developed by the ENCODE project [Gerstein, et al. 2012; Wang, et al. 2013a]. We found that regardless of the detection threshold (ranging from 1 in 100,000 fragment pairs called as peaks to 1 in 100), the pairs of interacting fragments called are significantly enriched with TFBS and active TSS (χ^2 p-value < 10^{-238}) with ~42% identified peak pairs overlapping with TFBS while ~28% identified non-peak pairs overlapping with TFBS (Figure 3.5). In addition, we found that fragments overlapping with TFBS or active TSS are involved in a slightly (but statistically significant) larger number of interactions than those not overlapping with TFBS or active TSS (Figure 3.6 and Figure 3.7).

Finally, we applied our methods to two other datasets: the mouse embryonic stem cell (mESC) and human embryonic stem cell (H1-hESC) dataset [Dixon, et al. 2012]. For both datasets, we downloaded observed Hi-C count data from the Fit-Hi-C website and estimated the expected counts using Fit-Hi-C. For mESC, genome-wide ChIA-PET data are available [Zhang, et al. 2013b] and for H1-hESC, 5C data were generated in 44 regions by the ENCODE pilot project [Sanyal, et al. 2012]. We gauged performance of our methods and of Fit-Hi-C by comparison with results from data generated from independent technologies. Comparison results suggest our method was able to detect more interactions captured from 5C (Figure 3.8) or ChIA-PET (Figure 3.9) data with the same number of peak regions called, according to HMRF posterior peak probabilities or Fit-Hi-C p-values.

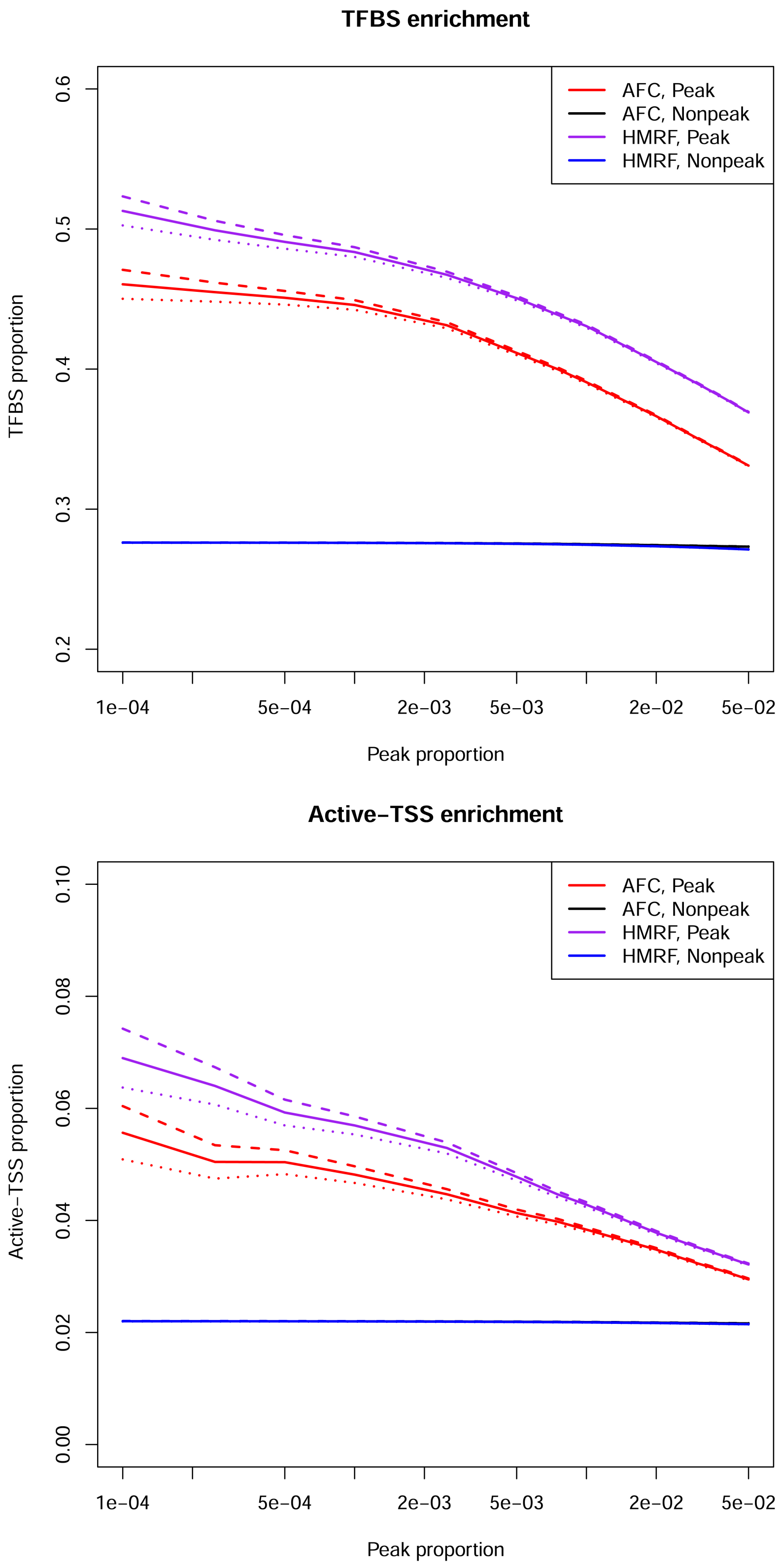


Figure 3.5 TFBS and active-TSS enrichment. Solid lines are estimated average levels. The 95% confidence interval of estimated average levels are represented by dashed lines and solid lines. Left panel: TFBS enrichment analysis. Right panel: active-TSS enrichment analysis.

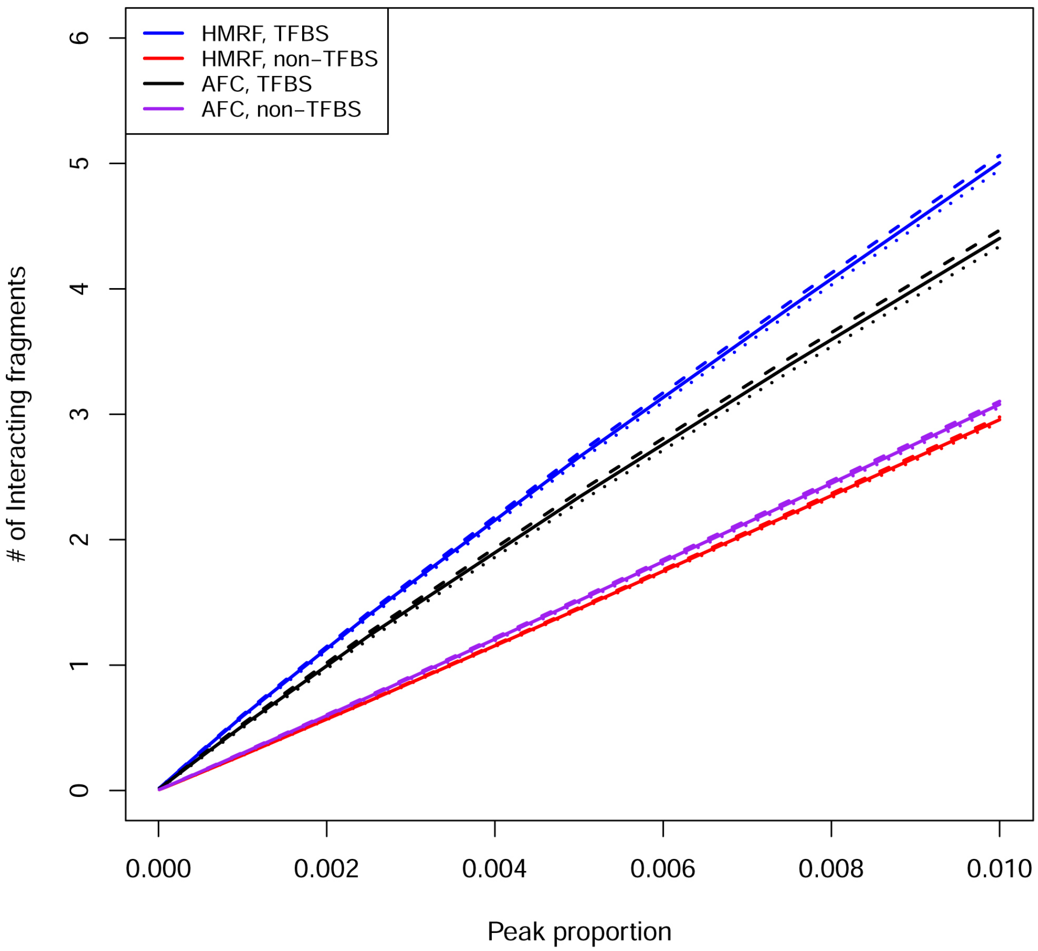


Figure 3.6 Transcription factor binding site (TFBS) enrichment analysis, at fragment level. The 95% confidence intervals are represented by dashed lines and dotted lines.

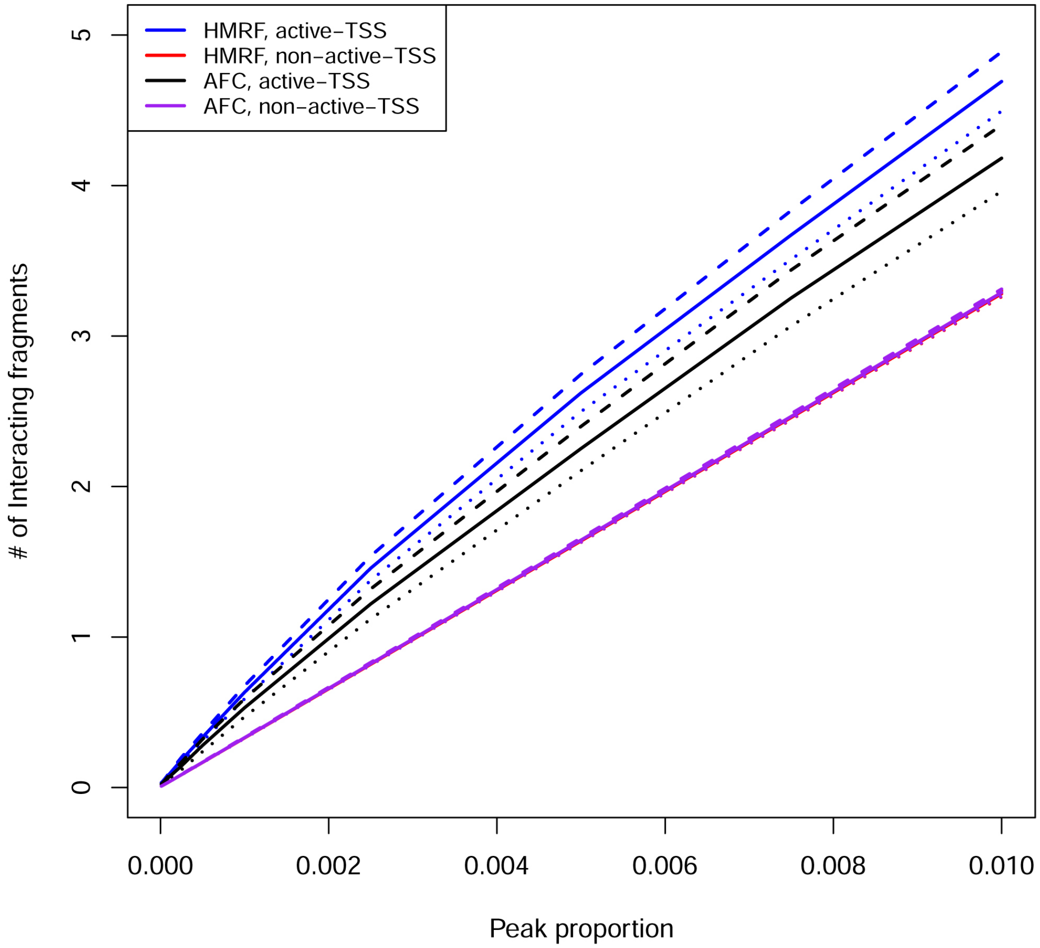


Figure 3.7 Active transcription starting site (TSS) enrichment analysis, at fragment level. The 95% confidence intervals are represented by dashed lines and dotted lines.

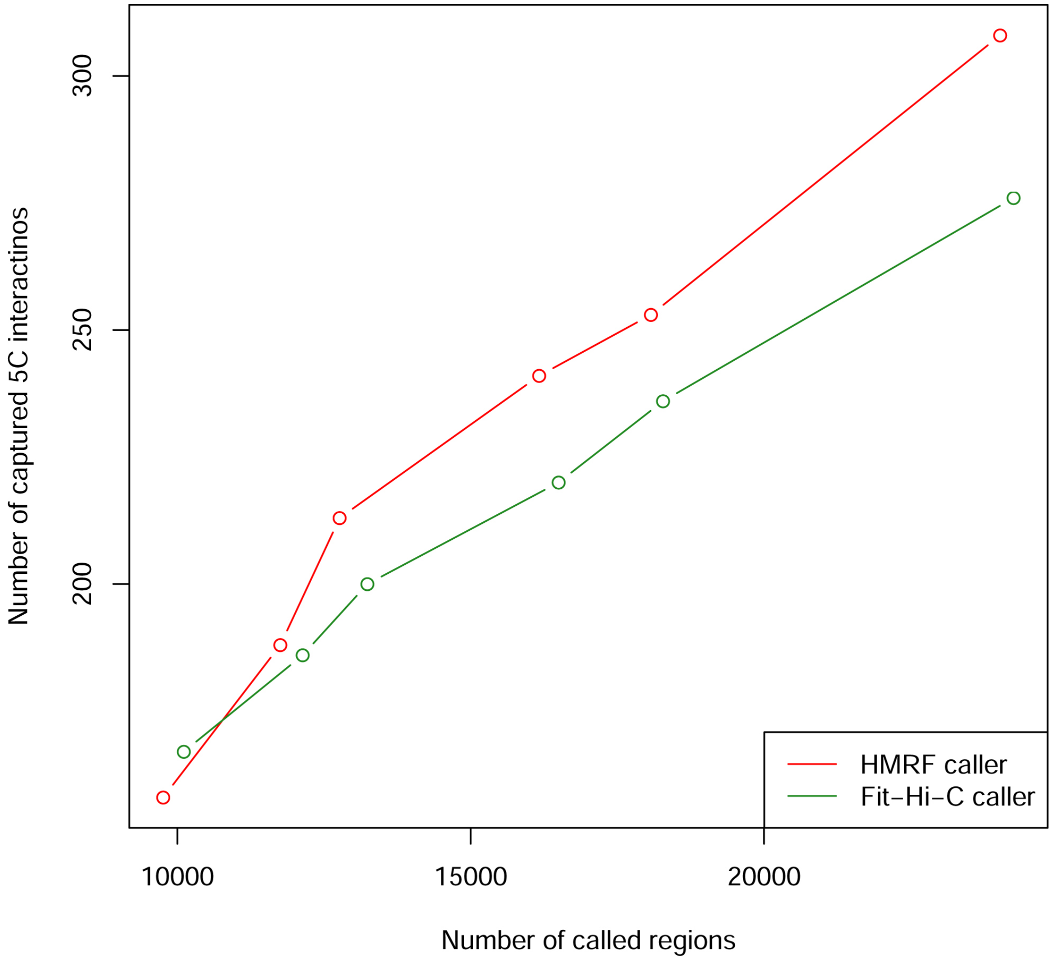


Figure 3.8 Power to detect interactions captured by hESC 5C dataset.

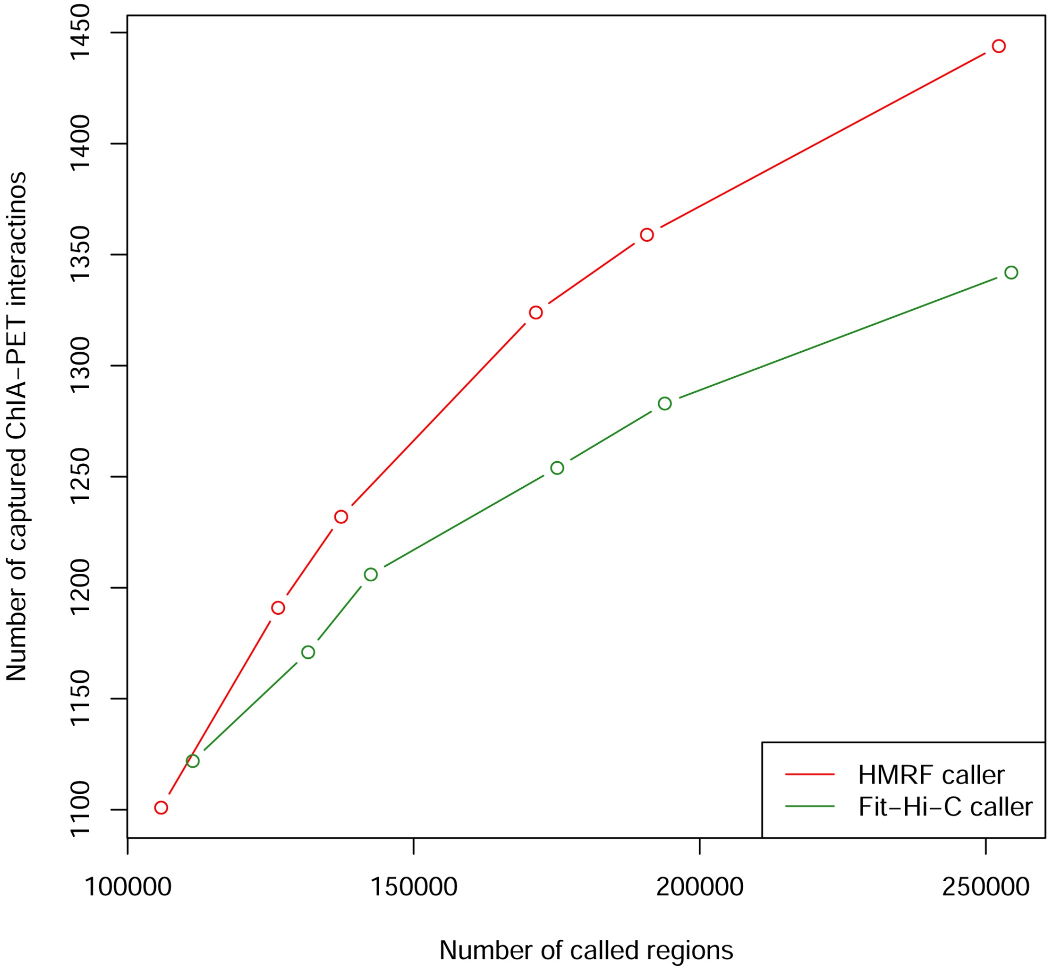


Figure 3.9 Power to detect interactions captured by mESC ChIA-PET dataset.

3.4 Discussion

Peak calling from data generated by 3C-derived methods is a fundamental task for the identification of chromatin interactions in 3D space. However, model-based methods for this important task are still lacking. Existing methods focus on the calibration of expected count frequency distribution under random collision, accounting for multiple biases behind 3C analysis including but not limited to density of restriction enzyme sites, mappability, and GC content. We have found existing methods rather mature for the purpose of calibrating expected counts (with results robust to different methods used). Establishing the expected count distribution is nevertheless a prerequisite, not peak calling itself. None of the existing methods consider the dependency underlying the peak status with statistical rigor.

In this work, we propose a HMRF based Bayesian method that explicitly models the dependency of the underlying peak pattern. The true peak pattern is unknown and can take different forms in the presence of dependency. We simplify the problem by assuming an Ising distribution prior and learn the level of dependency from data in a Bayesian framework. Our extensive simulations indicate superior performance in terms of both the estimation of the extent of dependency and the statistical power to distinguish peaks from background, across a range of underlying dependency patterns.

There are several aspects where the model can be further elaborated. First, our model has one θ, one ψ, and one ϕ, thus assuming that peaks are of similar strength and clustering patterns, and that reads have similar levels of over-dispersion. While the first two are simplifying assumptions bypassing issues including model selection difficulty and parameter non-identifiability, the last has been shown to be reasonable [Jin, et al. 2013]. Sensitivity analysis with the IMR90 combined dataset suggests these assumptions are reasonable: splitting each domain into two equal sub-domains resulted in highly consistent peak calls (Spearman correlation > 0.9). Second, we use a one-parameter Ising prior, with the parameter controlling both the peak proportion and level of dependency. A two-parameter Ising prior would allow more flexibility, particularly when the underlying dependency is weak. Third, our method could allow incorporation of prior knowledge, when available, into the model. For instance, a hyper prior could be imposed on the inverse temperature parameter based on estimated distribution from similar existing datasets. Finally, the computational complexity of our Bayesian modeling is quadratic in terms of the number of fragments under consideration. Our JAVA implementation takes ~13 minutes for a typical domain with 200 fragments and with parallel computing, genome-wide analysis can be easily accomplished within a few hours. In contrast, Fit-Hi-C and our *R* implementation of AFC take ~4 seconds and ~12 minutes, respectively. Therefore, for future work, computationally more efficient algorithms warrant consideration. We attempted to apply the iterative conditional mode algorithm [Li, et al. 2010], but observed unsatisfactory performance with weak peak signals (data not shown).

Despite these possible further model improvements, our method has demonstrated favorable performance over existing methods by borrowing information from neighboring fragment pairs via statistically modeling the potential dependency among the underlying peaks using a Bayesian framework. Our extensive simulation studies show the advantage of our method across a range of dependency patterns and its ability to learn the level of dependency (as modeled by the inverse temperature parameter) from data. Both are valuable since we have limited, if any, prior knowledge regarding the extent of dependency in real data. Re-analysis of several published Hi-C datasets including the IMR90, H1-hESC and mESC data confirmed the value of dependency modeling as taking dependency into consideration resulted in better concordance (> 40% improvement as measured by Jaccard Index of peak sets across two IMR90 datasets) and lower false positive, false discovery rates and higher recovery rate. Our method is the first to model dependency in a statistically rigorous manner and to borrow information from neighboring fragment pairs through a probabilistic model, was able to call up to 11.3% more enhancer-TSS interactions than the ad hoc method given the same number of peak regions called. We acknowledge that there is currently no genome-wide gold standard for real data (for example, from large scale genome-wide imaging-based experiments). We therefore made special efforts to benchmark the methods across multiple datasets and for each dataset, against the most reasonable silver standard. For example, for the IMR90 cell lines, we compared against results from the combined dataset with the highest sequencing depth, for H1 hESC and mESC, we used results from independent technologies (5C and ChIA-PET respectively).

With the continuing drop in sequencing costs and the intensive interest in chromatin structure as a way to understand GWAS results, we anticipate in the near future more high-resolution (fragment-level) Hi-C data where our method have demonstrated key advantage given the non-negligible dependency structure.

CHAPTER 4 ACROSS-PLATFORM IMPUTATION OF DNA METHYLATION

4.1 Introduction

Recently, the emergence of powerful technologies such as microarray-based DNA methylation studies [Bibikova, et al. 2011] and whole-genome bisulfite sequencing [Harris, et al. 2010] has enabled the profiling of DNA methylation levels at high resolution. Numerous studies employed these high-throughput approaches to characterize changes in DNA methylation patterns and their corresponding tissue and disease-specific differentially methylated regions on a genome-wide scale [Berman, et al. 2012; Chen, et al. 2014; Horvath 2013; Varley, et al. 2013].

As new technologies emerge, researchers tend to replace older methylation profiling platforms with new ones. However, different platforms can target CpG sites at different locations and with varying resolutions, which hinders the joint analysis of data from multiple platforms. For instance, the Illumina HumanMethylation27 (HM27) and HumanMethylation450 (HM450) BeadChip [Bibikova, et al. 2011] are two common microarrays used by The Cancer Genome Atlas (TCGA) project. While HM27 investigates 27,578 CpG sites predominantly located near CpG islands, HM450 provides broader coverage with 485,577 probes spanning 96% of CpG islands and 92% of CpG shores across a larger number of genes [Bibikova, et al. 2011]. Several TCGA studies have used HM450 to generate methylation profile data for more recently collected samples while still using HM27 to measure DNA methylation in the older test subjects. These mixed profiles compel researchers to focus on those probes shared between the two platforms when using the data for downstream analysis, as re-evaluating all samples using HM450 is not only expensive but also time-consuming [Cancer Genome Atlas Research 2013; Getz, et al. 2013; Koboldt, et al. 2012; Network 2012].

Imputation has been successfully employed in many genetic, genomic and epigenomic contexts [Donner, et al. 2012; Ernst and Kellis 2015; Jewett, et al. 2012; Li, et al. 2009; Zhang, et al. 2015]. However, no cross-platform imputation methods have been proposed for predicting methylation levels at unassayed CpG sites. On the other hand, for genotypes, imputation of untyped SNPs has become a standard procedure used both to resolve similar inconsistencies between genotyping arrays and to increase the resolution of genotype data collected in genome-wide association studies [Li, et al. 2009]. Here we propose the application of a similar concept to impute data in DNA methylation profiles from a subset of probes. Although DNA methylation does not exhibit as clear or strong a correlation structure as LD blocks among SNPs, we observe local correlation among neighboring probes similar as reported by others [Eckhardt, et al. 2006; Zhang, et al. 2015]. Importantly, we have found non-local correlations among probes falling into the same functional categories that have not been employed in the literature. Therefore we adopt a penalized functional regression model [Goldsmith, et al. 2011], which uses functional predictors to capture these non-local correlations. Our study demonstrates that this model can impute an HM27 dataset into an HM450 dataset effectively and accurately, and using these imputed values can improve the statistical power of downstream epigenome-wide association study (EWAS).

4.2 Materials and Methods

4.2.1 Data

We evaluated our imputation model using DNA methylation data from TCGA acute myeloid leukemia (AML) samples [Ley, et al. 2013]. The dataset contains DNA methylation data of tumor tissues from 194 patients with AML and is one of the largest methylation datasets from the TCGA project. All samples were evaluated using both HM27 and HM450. We transformed the raw β values into *M* values, defined as $M = \log_2[\beta / (1 - \beta)]$, as the *M* values better follow a Gaussian distribution [Cancer Genome Atlas Research 2013]. Our goal is to impute the HM27 dataset into an HM450 dataset to get an expanded view of the epigenomic landscape. The dataset is publicly available at the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/).

Since imputation of sporadic missing data is not the focus of this work, we removed all probes with at least one missing values for the sake of convenience. However, these missing values can be imputed by applying similar methods developed for gene expression profiles [Bo, et al. 2004; Kim, et al. 2005; Liew, et al. 2011; Troyanskaya, et al. 2001] to generate data without missing values. Additionally, we removed 743 probes designed in HM27 but not in HM450. In total, the HM27 dataset consisted of 20,794 probes passing TCGA quality control (QC) criteria [Ley, et al. 2013] and the HM450 dataset consisted of 393,152 QC+ probes. The latter set contained all 20,794 probes in HM27, leaving the remaining 373,358 as our potential imputation targets.

When training and using our model, we required data from HM450 and HM27, respectively. However, we noted that as HM27 and HM450 employ different biochemical methods to measure methylation levels, platform-specific effects might negatively impact imputation performance. To alleviate this systematic effect, we fitted a LOESS (locally weighted scatterplot smoothing) regression model [Cleveland 1979] between two platforms, stratified by the number of CpGs in the probe (#CpG = 0,1,2,3,4,5,6,7+), using 14 randomly chosen samples and normalized the HM27 data against the HM450 data [Cancer Genome Atlas Research 2013].

4.2.2 Penalized Functional Regression Model

We employed the penalized functional regression model [Goldsmith, et al. 2011] with minor modifications detailed below to quantify the relationship between DNA methylation from HM450 probes and DNA methylation density function estimated from HM27 probes together with other covariates. Specifically, assume for each target HM450 probe we have *n* observations and for each sample *i* = 1, 2, ..., *n*, we have data $\left[Y_i, X_i(t), Z_i\right]$, where Y_i is the transformed DNA methylation level at the target HM450 probe, $X_i(t)$ is the sample specific density function of the DNA methylation level measured by HM27 probes, denoted as T_i, and Z_i is a *p*-dimensional vector of covariates. We consider a functional linear regression model:

$$Y_i = \alpha + \int_0^1 X_i(t)\beta(t)dt + Z_i\gamma + \varepsilon_i$$

Here, α is the overall mean, $\beta(t)$ is the functional coefficient that characterizes the effect of density function $X_i(t)$ when $T_i = t$, γ is the regression coefficient vector for covariates, and $\varepsilon_i \sim N(0, \sigma^2)$.

To improve imputation accuracy, we incorporated functional predictors $X_i(t)$ into our model to capture information such as non-linear relationships from non-local probes. Based on the assumption that probes with similar properties tend to show similar methylation profiles, we divided the probes into several property groups. Here we divided the probes among five groups according to their relative location to a CpG island. The five groups are “CpG Island”, “North Shore”, “South Shore”, “North Shelf”, and “South Shelf” [Bibikova, et al. 2011]. Then we estimated the DNA methylation function $X_i(t)$ for a particular target probe with the DNA methylation data from HM27 probes in the same group as the target probe. Assume the target probe is in group *g* and there are *q* HM27 probes in the same group. The observed DNA methylation data are denoted as $t_j^g = \left(t_1^g, \dots, t_q^g\right)$, where t_j^g is the DNA methylation value at *j*-th HM27 probe in group *g* and *j* = 1, …, *q*. Instead of estimating $X_i(t)$ by expanding into the principal component basis obtained from its covariance matrix [Goldsmith, et al. 2011], we used the kernel density estimation to obtain $X_i(t)$ with t_i^g so that it is specific to group *g*.

To perform the model fitting, the functional coefficient $\beta(t)$ was expanded by a linear spline basis $\beta(t) = b_1 + b_2 t + \sum_{k=3}^{K_b} b_k (t - \delta_k)_+$, where δ_k are knots along the interval [0,1] and $(t - \delta_k)_+$ is an indicator function, taking value of 1 if $t > \delta_k$ and 0 if $t \le \delta_k$. We further defined a spline basis vector $\varphi(t) = \left\{\varphi_1(t), \varphi_2(t), \dots, \varphi_{K_b}(t)\right\} = \left\{\mathbf{1}, t, (t - \delta_3)_+, \dots, \left(t - \delta_{K_b}\right)_+\right\}$ and a coefficient vector $b = \left(b_1, \dots, b_{K_b}\right)'$ so that we may induce smoothing by assuming $b \sim N(0, D)$, where *D* is a penalty matrix corresponding to the particular spline basis $\phi(t)$.

Finally, we had $\int_0^1 X_i(t)\beta(t)dt = \int_0^1 f_{T_i}(t)\phi(t)bdt = \int_0^1 f_{T_i}(t)\phi(t)dt \cdot b$. For ease of notation, we denoted $J_{X\phi}$ as the $n \times K_b$ matrix with the (*i,k*)-th entry equal to $\int_0^1 f_{T_i}(t)\phi_k(t)dt$ and *Z* as the $n \times p$ matrix with the *i*-th row equal to Z_i*,* where *p* is the number of covariates. The model can be written in matrix format as

$$Y \middle| X(t) = \left[\mathbf{1}_n, J_{X\varphi}, Z\right]\left[\alpha', b', \gamma'\right]' + \varepsilon,$$

$$b \sim N(0, D).$$

This is a mixed effect model with K_b random effects *b* and penalty matrix

$$D = \begin{bmatrix} 0_{2\times 2} & 0_{2\times(K_b-2)} \\ 0_{(K_b-2)\times 2} & I_{(K_b-2)\times(K_b-2)} \end{bmatrix}$$

Typically, $K_b = 30$ is sufficient to avoid under smoothing in most applications [Goldsmith, et al. 2011]. Consistent with previous work [Fan, et al. 2015a; Fan, et al. 2015b], choice of K_b has little impact on performance (Figure 4.1 and Figure 4.2).

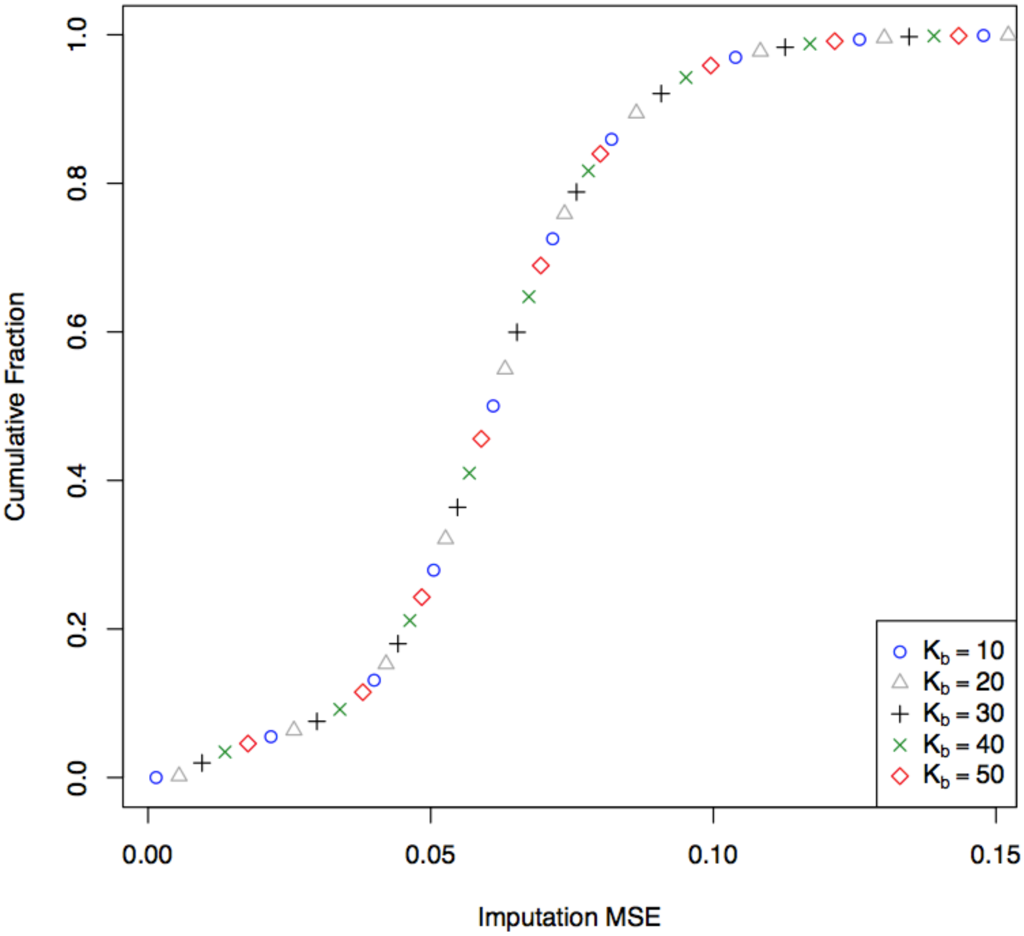


Figure 4.1 Empirical cumulative density function of imputation MSE for different K_b values. All lines overlap and are visually indistinguishable.

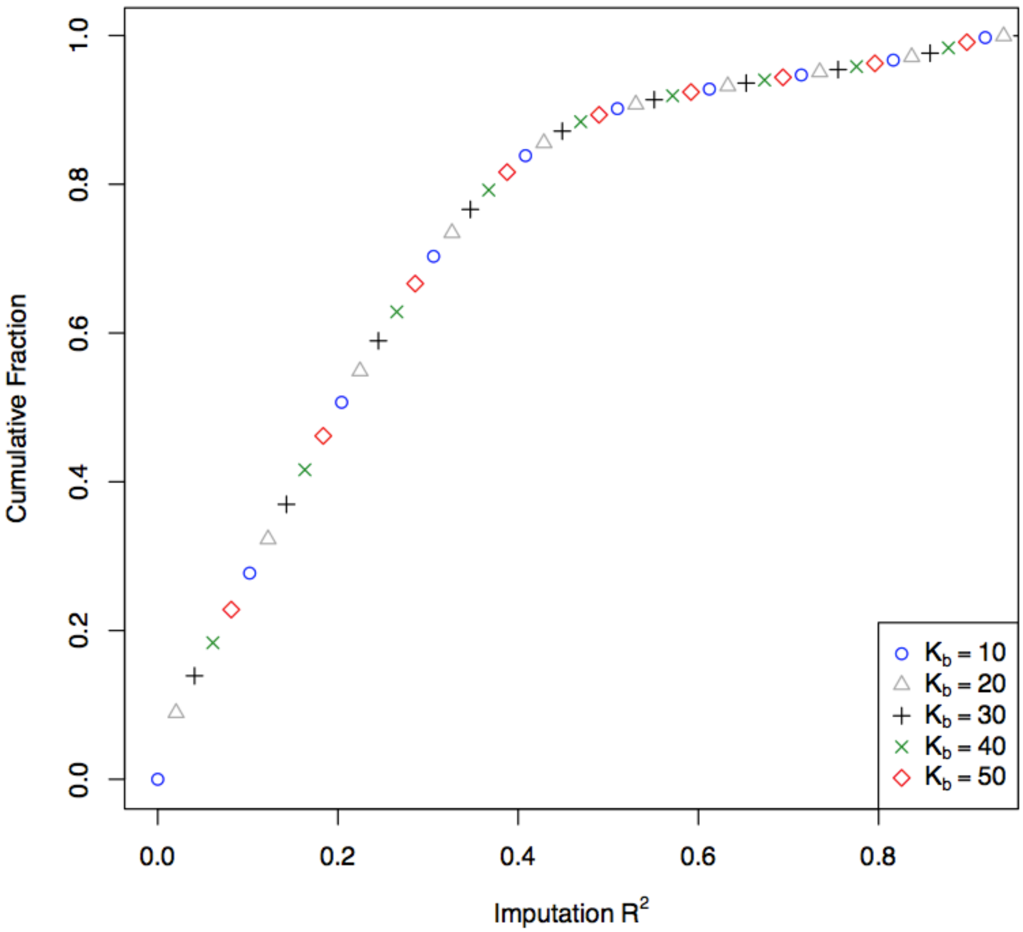


Figure 4.2 Empirical cumulative density function of imputation R^2 for different K_b values. All lines overlap and are visually indistinguishable.

4.2.3 Selection of Local Covariates

We exploited linear correlation with neighboring probes by including methylation values of HM27 probes near the target HM450 probe as local covariates Z in our imputation model. For simplicity, we selected the five nearest upstream probes and the five nearest downstream probes to each target probe as these local covariates.

4.2.4 Quality Filter

Since most probes showed nearly constant methylation levels across samples, we found for many probes, the imputation model is formed without sufficient information. Thus it tends to be under fitted and yields inaccurate imputation results. It is therefore desirable to have quality metrics for gauging the imputation quality. As such a quality metric, we proposed an under-dispersion measure defined as the ratio of the variance of fitted methylation values to its expected value (the variance of the true methylation values in the training set). If this ratio is below a certain threshold for a probe, it indicates an under fitted model for that probe, and we discard imputed values for the probe before subsequent analysis. A more stringent threshold can provide more accurate results, although at the cost of more probes discarded after imputation.

4.2.5 Imputation Quality Assessment

We assessed imputation quality using fivefold cross-validation. Within each split, the full dataset was randomly divided into a training set consisting of 80% of the samples and a testing set comprised of the remaining 20%. For each testing set, we only retained HM27 data, which contains a subset of HM450 probes, and masked methylation values of other HM450-specific probes. For the training set, we used methylation measurements on probes shared between the two arrays as predictors to impute methylation values at HM450-specific probes. Since most HM27 probes were measured by both HM27 and HM450, the predictors used in our model can be methylation levels for these shared probes measured from either array. Note that our prediction model was built under the realistic (more challenging) scenario where we used as predictors the measurements from HM450 array instead of those from HM27 array, which would require the training dataset had measurements from both arrays. Specifically, we fitted the functional regression model based on the training set, learned the relationship between methylation values of the shared and HM450-specific probes, and used the fitted model to impute the masked values of HM450 probes from the HM27 data in the testing set. Finally, we evaluated the imputation performance by averaging quality measures across splits.

As quality measures, we selected the mean squared error (MSE) and the squared Pearson correlation (R^2) between the imputed and the true methylation values in the testing sets. Although R^2 is a more intuitive measure of quality directly related to power and sample size in downstream analysis, we would like to note that this metric could easily be affected by a few outliers. Additionally, if the variance of methylation values for a specific probe is small, R^2 can be dramatically affected even by small imputation errors.

4.2.6 Simulation of Association Study

To assess the potential improvement of statistical power when using well-imputed methylation values for epigenetic association studies, we performed several simulated association studies for continuous and binary traits. Specifically, we randomly selected 100 HM450 probes with imputation R^2 between 0.1 and 0.3 based on our functional model, and simulated a dataset with 180 samples for each probe. In the continuous trait setting, for each probe, a trait value Y_i^* was simulated from the methylation level of this probe according to the linear model $Y_i^* = c\beta_i^* + \varepsilon_i$ for sample *i*, where β_i^* is true methylation β value, the effect size $c \in \{0, 0.1, 0.2, \ldots, 0.9, 1.0\}$, and $\varepsilon_i \sim N\left(\mathbf{0}, \mathbf{2}s_{\beta_i^*}\right)$, where $s_{\beta_i^*}$ is the sample standard deviation of β_i^*. In the binary trait setting, we first calculated $\eta_i^* = c(\beta_i^* - \bar{\beta}^*),\ p_i^* = \frac{e^{\eta_i^*}}{e^{\eta_i^*} + 1}$, and simulated Y_i^* from $Bernoulli(p_i^*)$, where $\bar{\beta}^*$ is the mean value of β_i^*, and the effect size $c \in \{0, 0.5, 1.0, \ldots, 4.5, 5.0\}$.

We repeated the simulation 2000 times. For each simulated dataset, we performed association tests (linear regression for continuous trait, and logistic regression for binary trait) based on the true methylation values, as well as imputed values from the simple linear model and our proposed penalized functional model. The empirical power of each method was calculated as the proportion of observed *p*-values that fall below the significance threshold $\alpha = 0.05$. Finally, we evaluated the empirical power for each effect size *c* by averaging results across 100 probes.

4.3 Results

4.3.1 Evaluation of Imputation Quality

Most probes showed nearly constant methylation levels in populations, making imputation trivial for them. We therefore focused on probes showing large variations and chose the top 20,000 such probes to evaluate the imputation quality. The time complexity of our method increases linearly with the number of target probes. However, since the imputation for each target probe is independent, we can accelerate it by running imputation in parallel. In the fivefold cross-validation experiment, 14 samples used for normalization were removed at first. Among the remaining 180, 144 individuals were chosen at random as the training set and 36 as the testing set within each split. The empirical cumulative distribution of imputation MSE and R^2 are shown in Figure 4.3 and Figure 4.4, respectively. The baseline method we used is the “tag” approach, where for each target probe, we calculated the Euclidean distance between the target probe and local probes, chose the local probe with the smallest distance as the tag probe and directly copied its methylation values as imputed values for the target probe. We also compared the two models with and without functional predictors and found that incorporating functional predictors leads to significantly improved imputation MSE and R^2 ($P<2.2 \times 10^{-16}$ for both metrics, paired Wilcoxon test). Table 4.1 summarizes some basic statistics. As expected, the “tag” method performs worst and we have therefore focused in subsequent text only the two models with and without functional predictors.

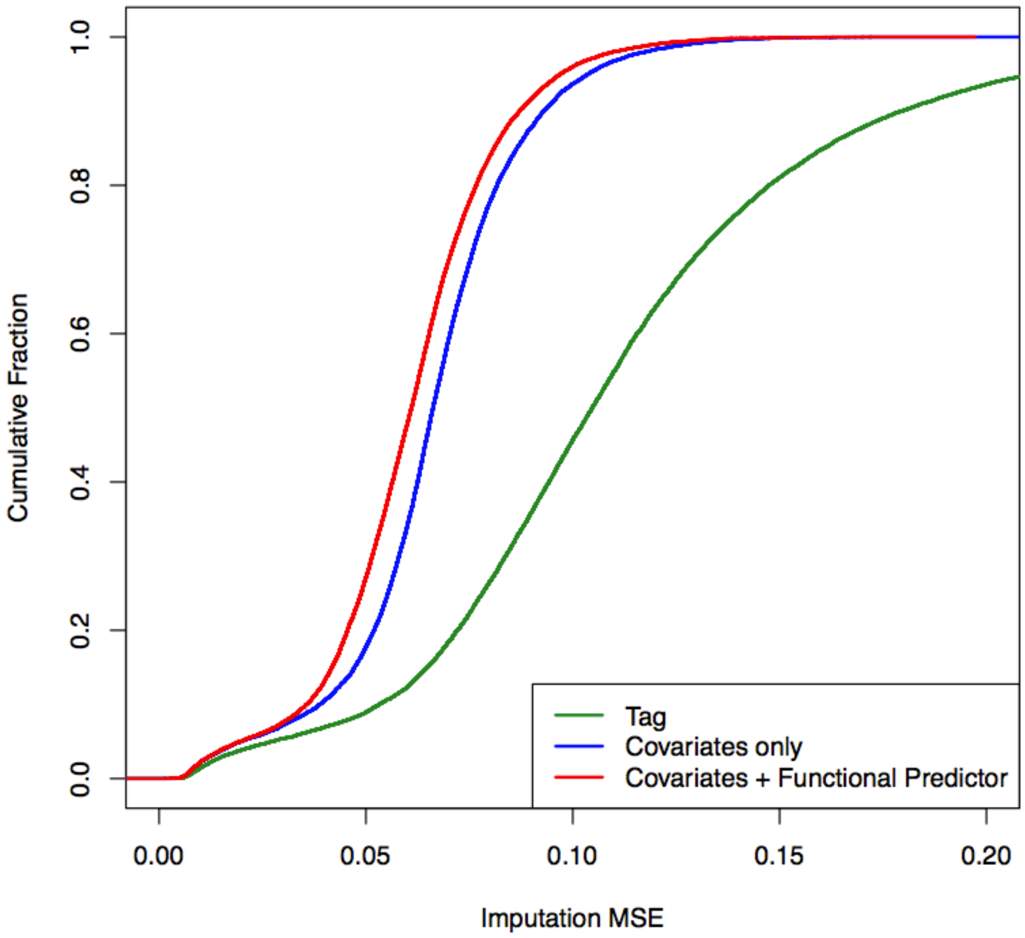


Figure 4.3 Empirical cumulative density function of imputation MSE for probes showing large variations in the AML dataset.

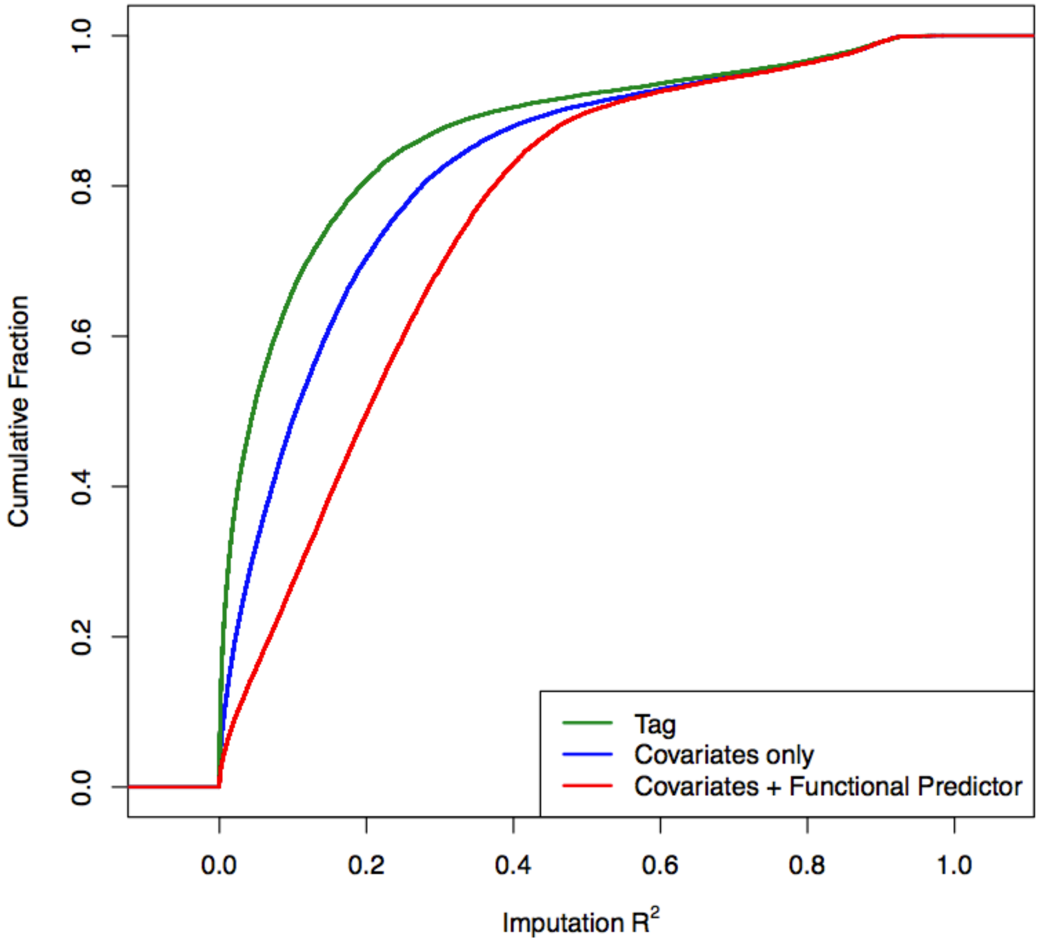


Figure 4.4 Empirical cumulative density function of imputation R^2 for probes showing large variations in the AML dataset.

Table 4.1 Quantiles of imputation MSE and R^2

	Imputation MSE			Imputation R^2		
	Q1	Median	Q3	Q1	Median	Q3
Covariates only	0.0553	0.0662	0.0781	0.0326	0.1040	0.2321
Covariates + Functional Predictor	0.0489	0.0610	0.0731	0.0907	0.2015	0.3375
Improvement	12%	8%	6%	178%	94%	45%

We used the target probe cg00288598 as an example to illustrate how the functional predictors improve the imputation quality. As shown in Figure 4.5, the selected local probes showed much smaller variation than the target probe, leading to an underfitted linear regression model and thus low imputation quality. In contrast, the methylation profile of the target probe is strongly associated with the distribution of methylation levels from all HM27 probes in its assigned North Shelf group, as indicated in Figure 4.6. Therefore after the functional predictors are added, the model can utilize the information from these non-local probes, including probes on different chromosomes, to alleviate the underfitting problem.

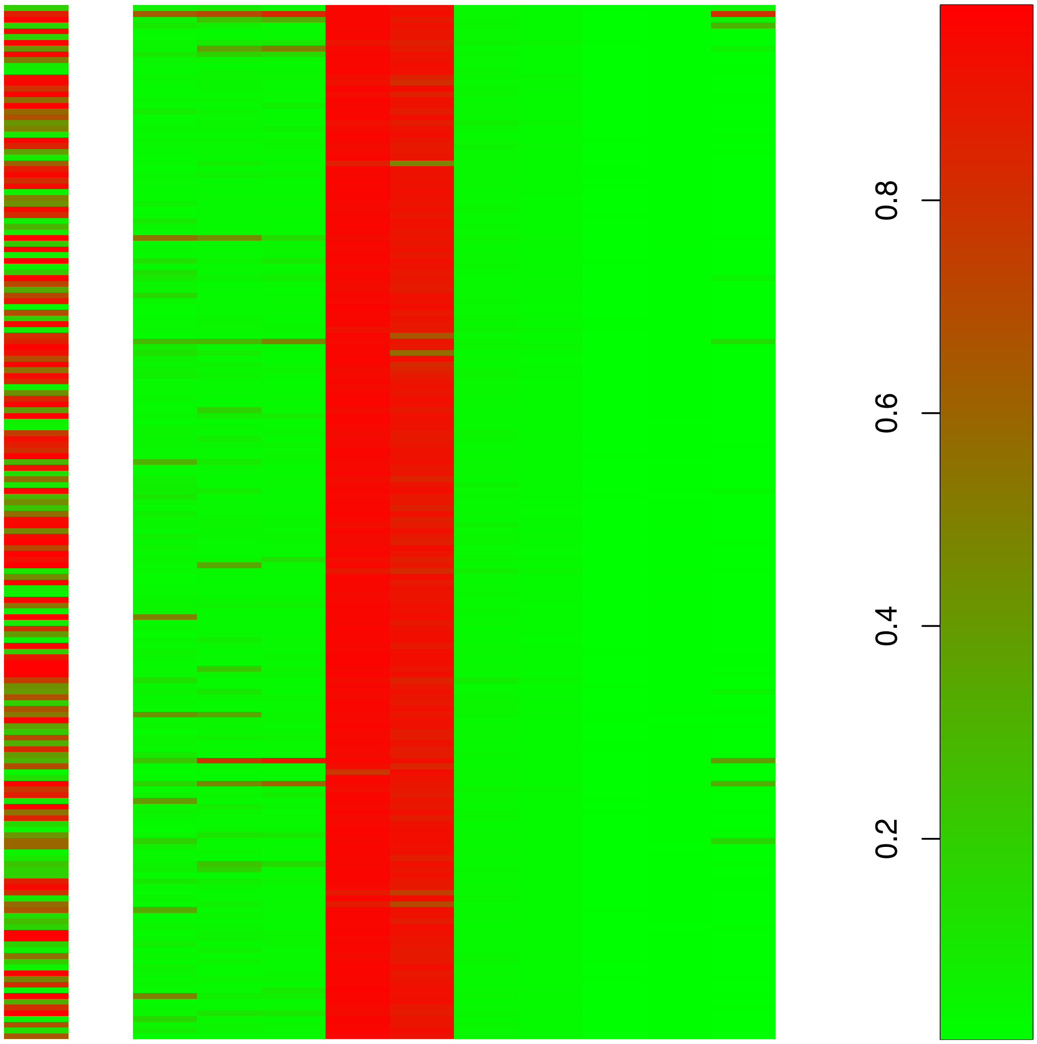


Figure 4.5 Methylation profiles of a North Shelf probe cg00288598 (left) and 10 selected local probes (middle).

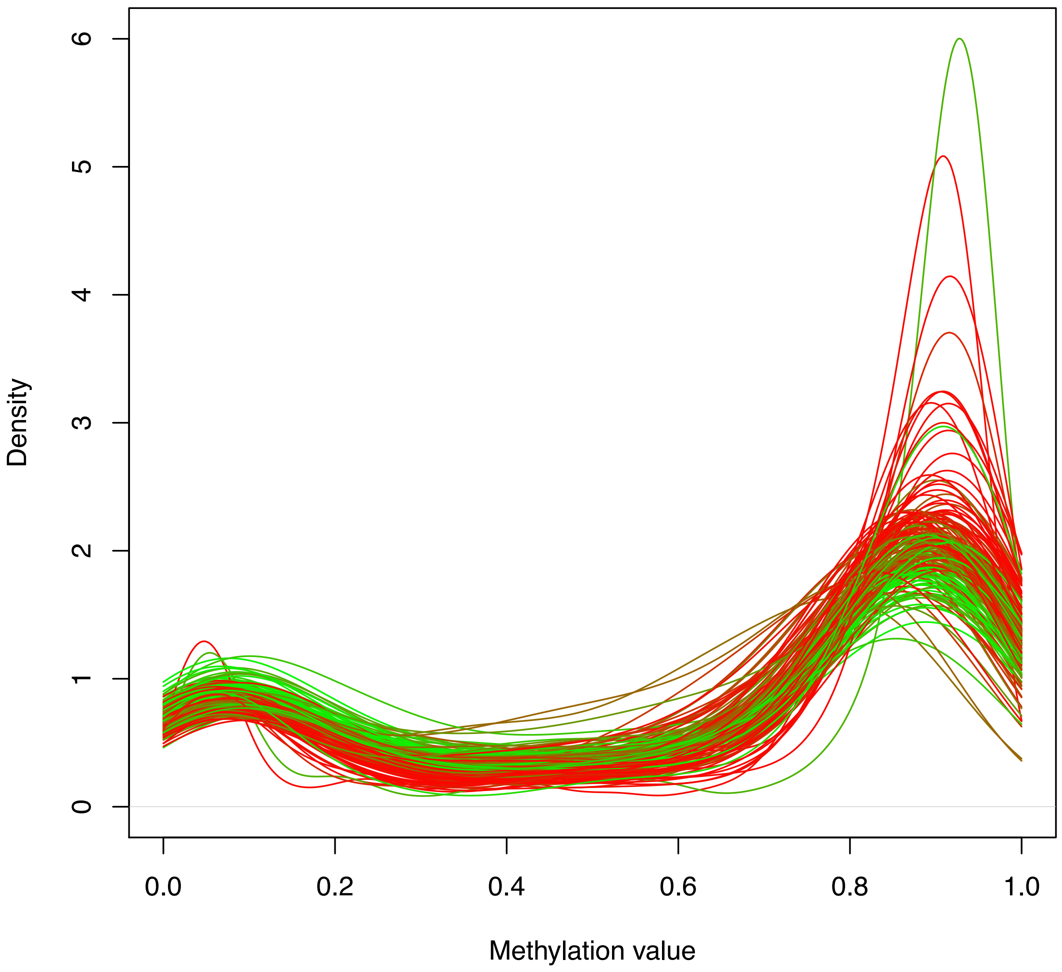


Figure 4.6 The individual-specific density plot of methylation values from all HM27 probes in North Shelf regions. Each line represents one individual and is colored based on the methylation level of the cg00288598 probe.

4.3.2 Performance of Quality Metrics

Because not all target probes can be imputed with the same level of accuracy, we tried to use the under-dispersion measure described in the Methods section to filter out inaccurate imputation results. We examined the relationship between imputation MSE/R^2 and the under-dispersion measure. We observed a negative correlation between the imputation MSE and this quality measure (Figure 4.7, Pearson correlation coefficient R = –0.65), and a positive correlation between imputation R^2 and the measure (Figure 4.8, Pearson correlation coefficient R

= 0.93). Therefore when performing imputation, we can calculate the under-dispersion measure and use it to filter out low-quality imputation results. Figure 4.7 and Figure 4.8 indicate that by choosing an appropriate threshold, we can remove most probes imputed with low-quality while simultaneously retaining nearly all probes imputed with high-quality. Based on our results, we suggest a threshold of 0.8 for the under-dispersion measure, which removes all badly imputed probes (defined as true $R^2 < 0.2$) at the cost of 1.24% well imputed probes (true $R^2 > 0.8$). Table 4.2 shows the number of probes passing post imputation quality filter at varying thresholds of the under-dispersion measure and we see that our penalized functional model results in up to 86.0% more probes that can be used for further analysis.

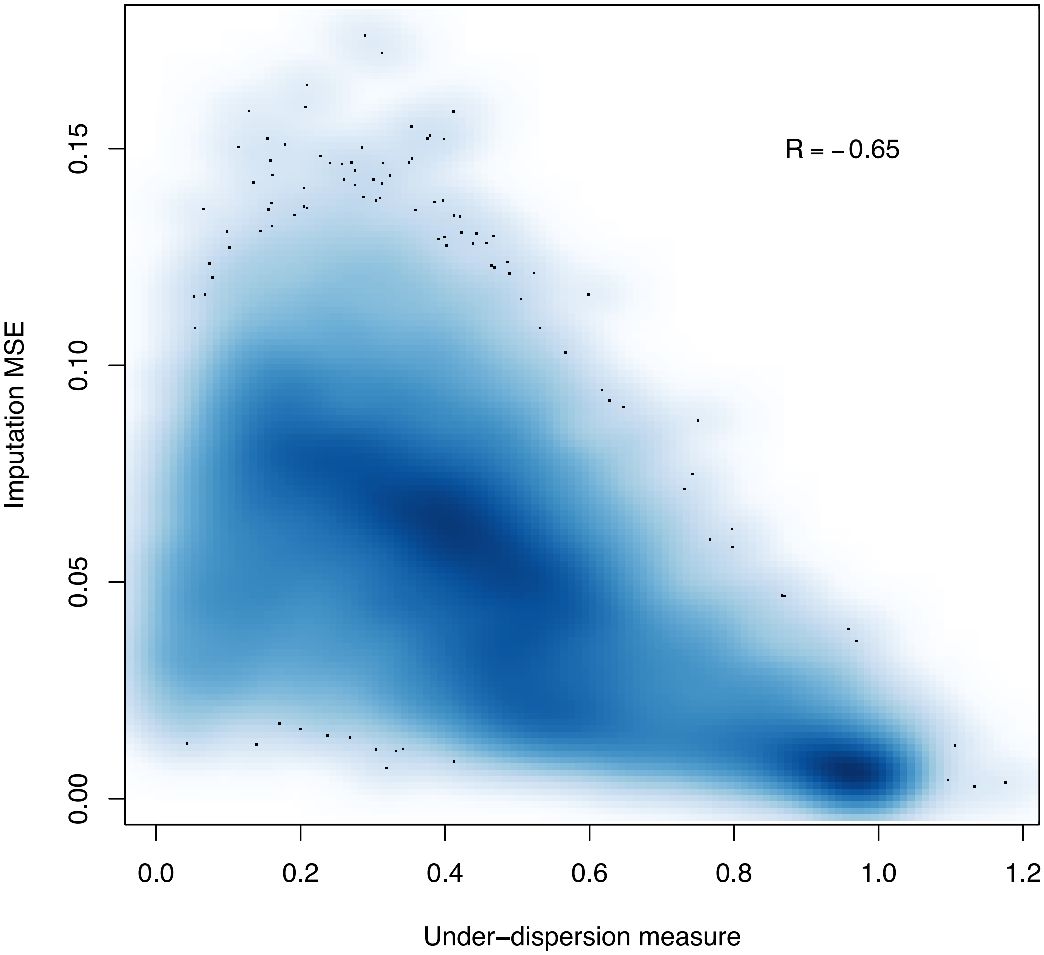


Figure 4.7 Scatter plot of under-dispersion measure and imputation MSE.

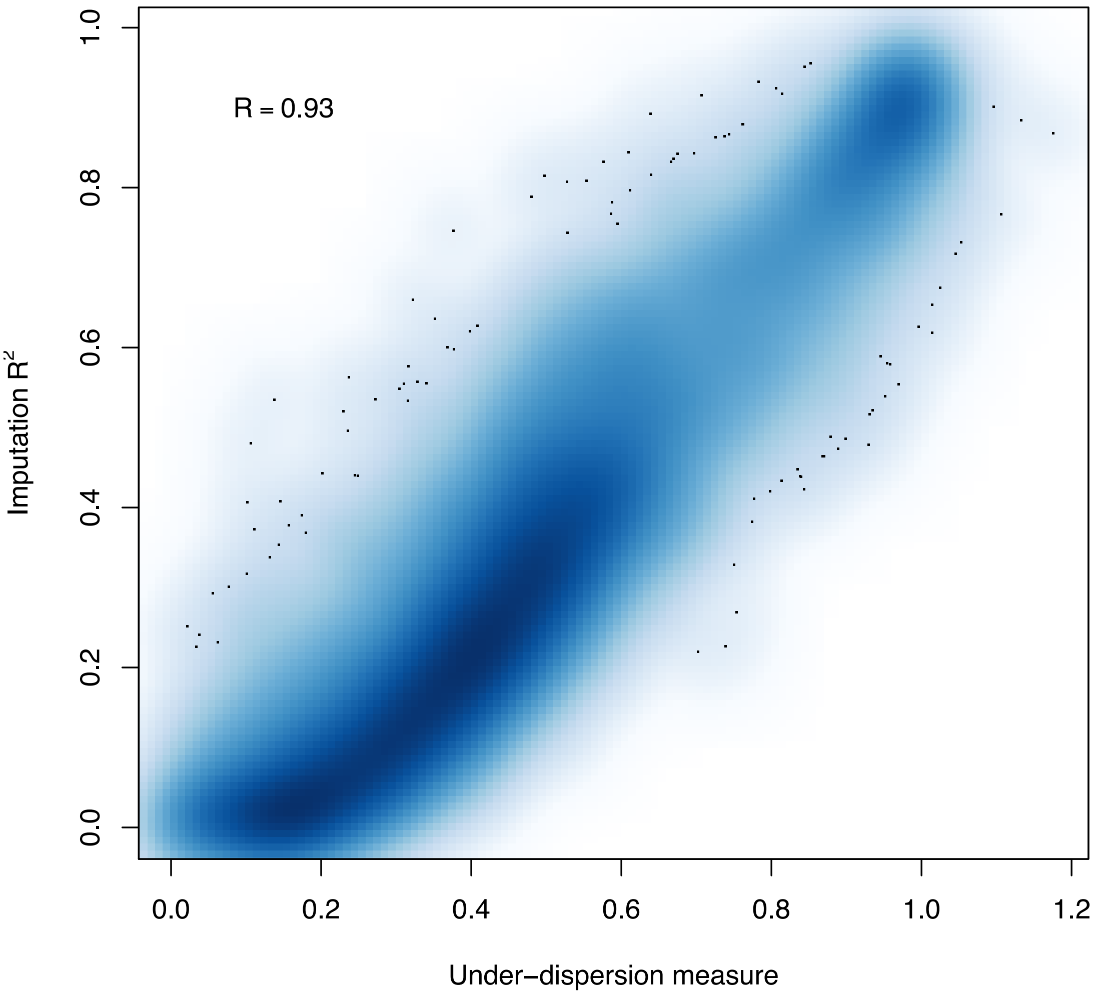


Figure 4.8 Scatter plot of under-dispersion measure and imputation R^2.

Table 4.2 Number of probes passing post-imputation quality filter

Under-dispersion measure threshold	0.6	0.7	0.8	0.9
Among top 20K probes				
Covariates only	2113	1592	1174	681
Covariates + Functional Predictor	2677	1691	1226	719
Improvement	26.7%	6.2%	4.4%	5.6%
Among all probes				
Covariates only	14479	8796	5123	2417

Covariates + Functional Predictor	26924	13117	6526	2684
Improvement	86.0%	49.1%	27.4%	11.1%

4.3.3 Power Gain in Association Study

It is not surprising to find relatively little difference in the performance of the two models at the two ends of the distribution (Figure 4.3 and Figure 4.4) because of probes that are either trivial or impossible to impute. Therefore in our work, we focus on the ~34% probes with imputation R^2 between 0.1 and 0.3 where our model demonstrates advantages over simpler models. As shown in Figure 4.9 and Figure 4.10, using imputed values from the penalized functional model for association tests is consistently more powerful than using values from the simple linear model, while the type I error rate (when $c = 0$) was still under proper control. These results suggest that even using probes with moderate imputation quality can substantially improve the statistical power of association test while maintaining the desired type I error rate.

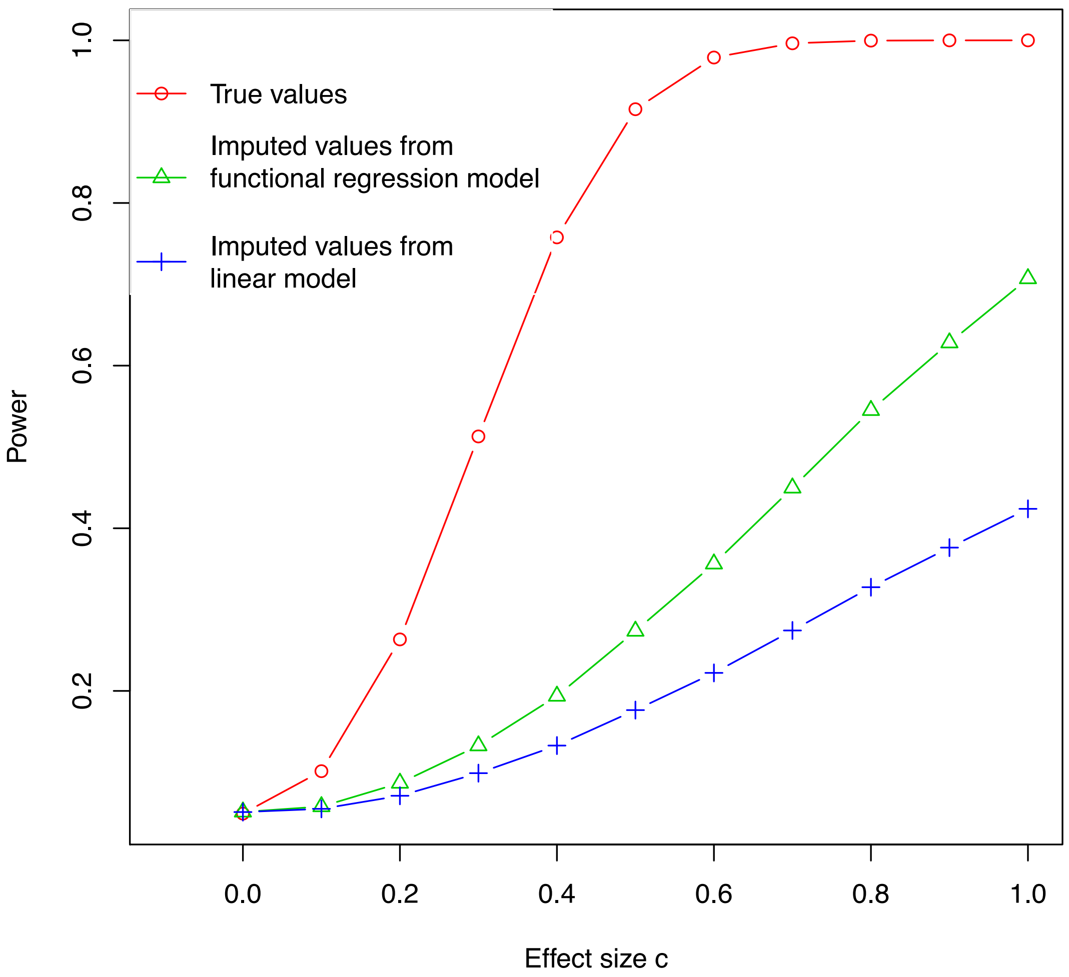


Figure 4.9 Empirical power of simulated association tests for continuous trait across a spectrum of effect size *c*.

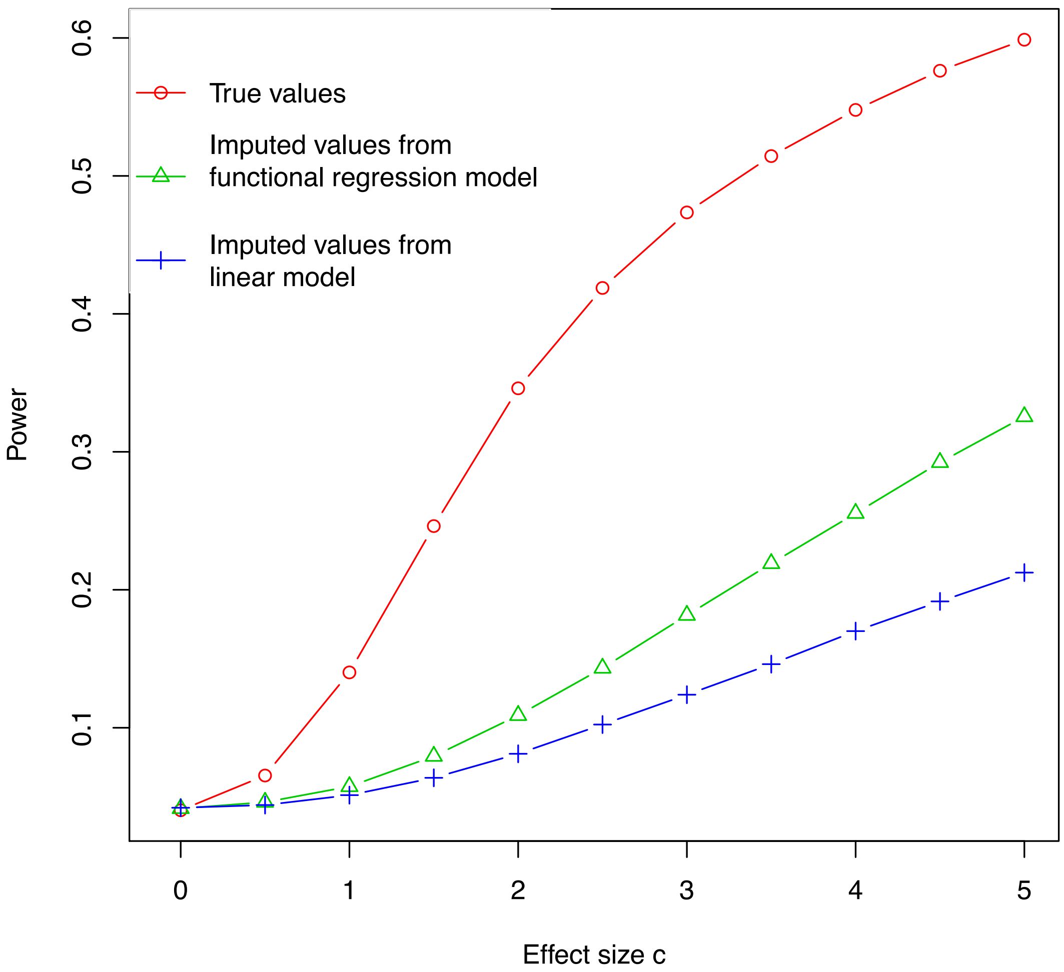


Figure 4.10 Empirical power of simulated association tests for binary trait across a spectrum of effect size *c*.

4.4 Discussion

In summary, we propose a penalized functional regression framework for across-platform imputation of methylation probes. Although a number of methods exist for predicting methylation levels at single CpG resolution, none of these directly apply to the across-platform imputation that we consider in this work. Moreover, we model information from non-local probes and have found such information considerably increase imputation performance. Our real data analysis demonstrates that by incorporating functional predictors from these non-local

probes, our model can produce accurate imputation results when the reference panel (training set) and target panel (testing set) characterize the same tissue under similar conditions.
Since DNA methylation profiles are highly tissue and condition-specific [Laurent, et al. 2010; Lister, et al. 2009; Varley, et al. 2013], our method will not work well if the two datasets are from different tissues or very different conditions. Recent studies suggest some statistical models to predict methylation profile in target tissue from a surrogate tissue [Ma, et al. 2014], which might be helpful in this case. Moreover, other systematic errors such as batch effect may also harm imputation quality. Therefore we suggest using techniques such as principal component analysis to check for obvious discrepancies between reference and target panels before applying our method.

In various settings, a different way to construct predictors may further improve the performance of our model. For example, non-local probes can be categorized based on other properties, such as their relative location to a gene [Bibikova, et al. 2011]. Another possible approach to select non-local probes is to choose HM27 probes highly correlated with the target probe (See supplementary methods). Supplementary Figure S1 shows that this approach can lead to better imputation performance, but the computational cost will be much higher. We can also explore other approaches to select local covariates, such as using a different number of probes, or choosing the local covariates as the 10 local probes that have the highest correlation with the target probe.

Since most CpG sites display stable DNA methylation levels, imputation error is low on average (the median imputation RMSE for beta values of all probes is ~0.05). Dichotomizing at beta value of 0.5 following Zhang et al [Zhang, et al. 2015], our prediction accuracy is 94.9%, largely consistent with their reported 92% prediction accuracy. However, researchers may consider dynamic CpG sites to be of more interest, as these sites often co-localize with key regulators such as enhancers and transcription factor binding sites [Ziller, et al. 2013]. Therefore we calculated quality metrics for individual probes, facilitating the evaluation of imputation quality for each probe and removing probes with low imputation quality for downstream analysis. For probes showing large variation of methylation levels, we notice that even after incorporating functional predictors, the imputation quality is still low for a significant portion of these probes. Possible reasons are: First, the DNA methylation profile alone does not provide sufficient information for accurate imputation. We may need to incorporate other information to improve imputation quality, such as local DNA context and the binding profile of regulatory proteins [Bhasin, et al. 2005; Bock, et al. 2006; Zheng, et al. 2013], although this requires additional data source in the same or similar tissue type that are rarely available. Second, HM27 has a much lower resolution than HM450. In addition, a large proportion of HM27 probes showed nearly constant methylation levels across samples. As such, an extreme case is that if the target HM450 probe is not correlated with any HM27 probes, the model will be under fitted with the predicted methylation levels for all samples close to the average, thus leading to smaller variance than expected, similar to under dispersion observed with imputed SNP data [Li, et al. 2009]. We expect to observe better performance if we impute from a denser microarray. Third, our normalization procedure does not fully eliminate the inconsistency of measurements between HM27 and HM450, which also affects the performance of our model. Here we assumed only HM450 data is available for the training dataset, which is a more realistic setting. However, if the training set contains both HM27 and HM450 data in a real case, we can treat HM450 data as response and use HM27 data to construct predictors. Thus predictors from both training and testing set are constructed from HM27 data and the inconsistency between HM27 and HM450 is automatically learned by the model. In this case, our model will show higher imputation accuracy.

Since a considerable proportion of CpG probes on HM450 overlap with SNPs (hereafter referred to as SNP-probes), we also examined whether imputation quality for these SNP-probes differ from that for non-SNP probes. Our annotation [Barfield, et al. 2014] includes 98,741 CpGs that have a SNP somewhere underneath the 50bp probe, among which 62,777 are QC+ HM450-specific sites. We found that the SNP-probes are slightly less varying than the non-SNP probes (for example, median variance of β values is 0.00310 and 0.00356 respectively) (Table 4.3). Analogous to rarer variants in SNP imputation [Duan, et al. 2013; Li, et al. 2009; Liu, et al. 2012; Pistis, et al. 2015], it is not surprising to find that these SNP-probes appear slightly easier to impute when measured using MSE (for example, median MSE is 0.00236 and 0.00263 respectively) but actually slightly more challenging to impute when measured using the more honest information content R^2 metric (median R^2 is 0.162 and 0.182 respectively).

Table 4.3 Imputation quality of SNP-probes versus non-SNP-probes

	Variance in β measurement		Imputation MSE		Imputation R^2	
	Mean	Median	Mean	Median	Mean	Median
SNP probe	0.0131	0.00310	0.0110	0.00236	0.206	0.162
Non-SNP probe	0.0140	0.00356	0.0115	0.00263	0.223	0.182

The focus of the present work is on imputation per se rather than association analysis. After accurate imputation, we can combine data from multiple platforms to obtain methylation levels of more CpG sites for downstream analysis such as detecting methylation quantitative trait loci or EWAS [Heyn and Esteller 2012; Rakyan, et al. 2011]. Such analysis can take imputation uncertainty into account similarly as for imputed SNPs [Huang, et al. 2014]. In this work, we evaluated statistical power under the mostly commonly observed change in mean values, however, other forms of changes have been observed. For example, several studies [Gervin, et al. 2011; Hansen, et al. 2011] reported differences in the variation (in addition to the mean) of methylation values between cancer and healthy groups. Our simulation studies show power improvement even using standard logistic regression to test mean difference under such variation difference. Regardless of the epigenetic architecture on phenotype, we expect our imputation method, resulting in higher-resolution and more powerful exploration of the epigenome, will lead to rapid advances in understanding the functional role of normal DNA methylation and the impact of its aberration. Our method is implemented in R and freely available at https://github.com/Leonardo0628/pfr.

CHAPTER 5 INTEGRATIVE ANALYSIS OF INFLAMMATION-RELATED GENES IN HBV-RELATED HEPATOCELLULAR CARCINOMA

5.1 Introduction

Hepatocellular carcinoma (HCC) is one of the most common cancers in the world. The major risk factors for the development of HCC include infection with hepatitis B virus (HBV), hepatitis C virus (HCV), aflatoxin exposure, and chronic alcohol abuse [El-Serag and Rudolph 2007]. It has been reported that about 55% of HCC occurs in China, where the major etiological factor is chronic HBV infection [Tanaka, et al. 2011]. HCC is one clear example of inflammation-related cancers, with chronic inflammation being indispensable in its development. It has been shown that chronic liver inflammation due to persistent HBV or HCV infection may lead to cirrhosis, which can eventually progress to HCC at an incidence rate that is 4~5 times higher than that among asymptomatic HBV carriers [Fattovich, et al. 2004].

During the past decades, evidence has been accumulated to show that dysregulation of inflammation-related genes plays important roles in the development of HCCs [Zitvogel, et al. 2012]. It was reported that both chromosomal aberrations such as copy number loss or gain and epigenome deregulation by DNA methylation, histone modification and non-coding RNAs may contribute to the aberrant expression of inflammation-related genes [Lee 2013; Natoli, et al. 2011; Ozen, et al. 2013; Zhang 2012]. However, current knowledge is mainly derived from incomplete studies focusing on only single or a few such genes, using either a candidate gene or genome-wide approach. To the best of our knowledge, there are few, if any, comprehensive studies employing integrative analysis to simultaneously interrogate both genetic and epigenetic events contributing to the aberrant expression of genes in the inflammation pathway in HBV-related HCCs.

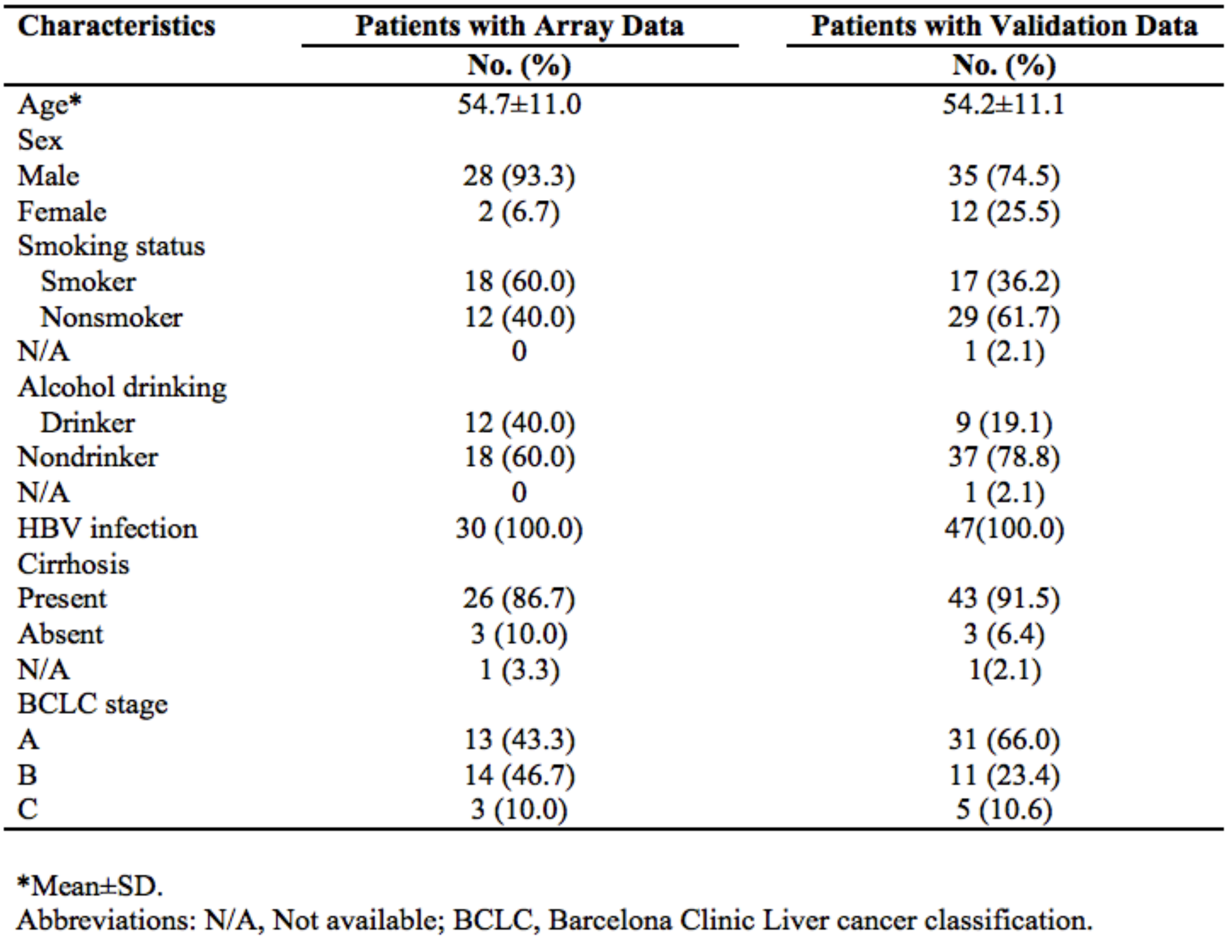
In this current study, we used high throughput array-based technology to comprehensively analyze the relationship between DNA methylation or somatic copy number aberration (SCNA) and aberrant expression of 1,027 genes in the inflammation pathway [Loza, et al. 2007] in HBV-related HCCs. We validated our array-based results in public datasets and in an additional sample set of HCCs and paired non-tumor tissues. Our data indicated that DNA methylation and SCNA indeed cause some inflammation-related genes to be aberrantly expressed, but they only contribute about 30% aberrant expression of these genes in HBV-related HCCs.

5.2 Materials and Methods

5.2.1 HCC Samples

In this study, we performed a two-stage analysis in total up to 77 HCCs and their paired non-tumor tissues (more than 2 cm from tumor). In brief, 30 HCCs and their paired non-tumor specimens were randomly selected for the discovery phase and the rest 47 HCCs and their paired non-tumor tissues were used for the validation phase. All tumor and their paired non-tumor tissues were obtained from hepatectomy of patients with HCC between 2010 and 2013 at Cancer Hospital, Chinese Academy of Medical Sciences (Beijing). The samples were immediately frozen in liquid nitrogen upon surgically resected. The diagnosis of HCC was all confirmed by histopathology. We selected the samples for HBV positive but HCV negative according to serology tests and infection history. Patients who had received chemotherapy or radiotherapy were excluded from this study. We also collected clinical characteristics of each subject in this study, which are shown in Table 5.1. All patients signed an informed consent and this study was approved by the Institutional Review Board of the Chinese Academy of Medical Sciences Cancer Institute.

Table 5.1 Clinical characteristics of 30 HCC patients with array data and 47 patients with validation data



Characteristics	Patients with Array Data	Patients with Validation Data
	No. (%)	No. (%)
Age*	54.7±11.0	54.2±11.1
Sex		
Male	28 (93.3)	35 (74.5)
Female	2 (6.7)	12 (25.5)
Smoking status		
Smoker	18 (60.0)	17 (36.2)
Nonsmoker	12 (40.0)	29 (61.7)
N/A	0	1 (2.1)
Alcohol drinking		
Drinker	12 (40.0)	9 (19.1)
Nondrinker	18 (60.0)	37 (78.8)
N/A	0	1 (2.1)
HBV infection	30 (100.0)	47(100.0)
Cirrhosis		
Present	26 (86.7)	43 (91.5)
Absent	3 (10.0)	3 (6.4)
N/A	1 (3.3)	1(2.1)
BCLC stage		
A	13 (43.3)	31 (66.0)
B	14 (46.7)	11 (23.4)
C	3 (10.0)	5 (10.6)

*Mean±SD.
Abbreviations: N/A, Not available; BCLC, Barcelona Clinic Liver cancer classification.

5.2.2 Arrays Used in the Discovery Phase

In the discovery phase, high-throughput screening was performed using Affymetrix Human Gene 1.0 ST Array (Affymetrix, Santa Clara, CA), Nimblegen 3×720 K CpG Island Plus RefSeq Promoter Array (Roche NimbleGen, Madison, WI) and Affymetrix GeneChip Human Mapping 6.0 array (Affymetrix) to measure mRNA expression, DNA methylation and copy number changes in HCCs and paired non-tumor specimens, respectively. Referring to the corresponding array design database, we confirmed that the Affymetrix Human Gene 1.0 ST Array interrogates all 1,027 inflammation-related genes described by Loza et al. with 1,108 transcripts; and the Nimblegen 3×720 K CpG Island Plus RefSeq Promoter Array and Affymetrix GeneChip Human Mapping 6.0 array interrogate 38,179 probes covering 1,024 genes and 33,855 probes covering 938 genes, respectively.

5.2.3 Array-based Data Production

Total RNA samples were extracted from 30 fresh HCCs and paired non-tumor tissues using the Trizol reagent (Life Technologies, Carlsbad, CA). Genome-wide transcriptional profiling was produced using Affymetrix Human Gene 1.0 ST Array according to the manufacturer's protocol. Arrays were processed in two batches; one included 10 arrays, and the other included 20 arrays. Raw data were first processed using Robust Multiarray Averaging [Bolstad, et al. 2003] and batch effect was adjusted using ComBat [Johnson, et al. 2007]. Adjusted expression values were used in subsequent downstream analysis.

Genomic DNA samples were isolated from the same 30 fresh HCCs and paired non-tumor specimens using a commercial DNeasy Blood & Tissue Kit (QIAGEN, Valencia, CA). Each DNA sample was then divided into two portions. One portion was bisulfite-converted using the EZ DNA Methylation kit (Zymo Research, Irvine, CA), and the DNA methylation profiles was obtained with the MeDIP-chip platform based on the Nimblegen 3×720 K CpG Island Plus RefSeq Promoter Array according to the manufacturer's protocol. Arrays were also processed in two batches in the same manner as afore-described for gene expression quantification. The pre-processing of DNA methylation array data was similar to that used for gene expression array data. We first normalized the raw data using control probes designed in the methylation array, and then adjusted batch effect with ComBat.

The other portion of genomic DNA sample was used to detect DNA copy number aberrations using the Affymetrix GeneChip Human Mapping 6.0 set according to the manufacturer's protocol. Affymetrix Power Tools was used on the raw data to generate signal intensities, which were further analyzed by PennCNV [Wang, et al. 2007] to call probe-based copy numbers. Log R ratios (LRR) estimated by PennCNV were carried forward for further analysis.

5.2.4 Array-based Data Analysis

Paired student's t-test was used to determine whether the difference in gene expression, DNA methylation or copy numbers of inflammation-related genes between HCC tissues and matched non-tumor liver tissues is significant. Bonferroni adjustment was used to correct for multiple comparisons in view of 1,108 transcripts, 38,179 methylation probes, and 33,855 SNP/CN probes covering inflammation-related genes; therefore, P values $< 4.5\times10^{-5}$, $< 1.3\times10^{-6}$ and $< 1.48\times10^{-6}$ were considered to be statistically significant, respectively, for transcripts, methylation and copy number. In this study, significant methylation probes covering the same gene locus all showed effect in the same direction, except for the 2 significant probes covering the FYN gene. Therefore, we calculated the mean methylation value of all significant probes across the same gene for further gene-level DNA methylation analysis. As to the FYN gene, the 2 significant probes were both taken forward for separate further analysis.

Because of tumor heterogeneity, the abundance and size of SCNA often vary across different patients with HCC or even across different tumor cells from the same tissue. Besides, although several analytical programs have been established to detect SCNAs based on the intensity of SNP array probes, the results obtained by using these programs are not all consistent. Therefore, in this study, we used a simple spanning strategy [Genomes Project, et al. 2012], where only DNA regions that contain at least 2 adjacent SNP/CN probes showing significant difference between tumor and paired normal tissue were considered as SCNA markers. For significant probes exhibiting changes in the same direction at one gene locus, gene level mean LRR value across the significant probes was used for further analysis. For probes exhibiting changes in opposite directions at the same gene locus, mean LRR values across the significant probes were separately calculated for each direction and carried forward for further analysis. Spearman correlation coefficient was used to test the correlation between gene expression and DNA methylation or copy numbers for each gene.

5.2.5 Validation Using Public Dataset

Public datasets, GSE14520 and GSE25097, or GSE37988 and GSE54503 datasets, together with our data, were used to identify the overlapping genes reported by array-based gene expression or DNA methylation. The processed datasets of GSE14520, GSE25097, GSE37988 and GSE54503 were extracted from GEO database (http://www.ncbi.nlm.nih.gov/geo/), and analyzed by GEO2R software. Student's t-test was used to examine the difference in aberrant expression or DNA methylation changes of inflammation-related genes in HCCs with Bonferroni corrected significance threshold. In addition, we validated the SCNA results in the GSE38323 dataset, and validated the correlation between gene expression and SCNA in the published GSE28127 dataset [Lamb, et al. 2011; Wang, et al. 2013b].

5.2.6 Validation in the Independent Sample

An independent sample, consisting of 47 surgically removed HCC and paired non-tumor specimens, was used to further validate our findings from the discovery stage.

Total RNA isolated from each tissue specimen was converted to cDNA using oligo(dT)15 primer and SuperScriptII (Invitrogen, Grand Island, NY). mRNA levels were measured by quantitative real-time PCR (RT-PCR) on an ABI Prism 7900 sequence detection system (Applied Biosystems, Foster City, CA) using the SYBR Green method. mRNA levels of the candidate genes were calculated relative to expression of GAPDH.

Methylation profile was determined in genomic DNA isolated from each tissue specimen. Primers targeting the promoter regions or CpG islands of the candidate inflammation-related genes were designed as described by Wojdacz et al. [Wojdacz, et al. 2008]. Tissue DNA samples, commercial fully methylated DNA and unmethylated DNA (Zymo Research) were simultaneously converted by bisulfite. A series of methylation dilution standards of 100%, 75%, 50%, 25%, 10%, 5% and 0% were prepared to depict methylation standard curve by mixing methylated DNA and unmethylated DNA. RT-PCR and methylation-sensitive high resolution melting (MS-HRM) analysis were carried out on an ABI Prism 7900 sequence detection system.

RT-PCR based on SYBR Green method was used to examine copy numbers of the candidate genes in each genomic DNA sample, with the LINE-1 gene, the most abundant retrotransposon in the human genome as the reference [Wang, et al. 2002]. Copy numbers of the candidate genes were calculated relative to those of LINE-1 in each sample.

Paired student's t-test was used to examine the differences in gene expression, DNA methylation and copy numbers, at a false discovery rate (FDR) of 0.05.

5.2.7 Gene Network Construction

Inflammation-related genes with expression changes tallying with DNA methylation changes or copy number changes were selected to construct gene networks using the MetaCore database and software (GeneGo, Inc.; http://thomsonreuters.com/metacore/).

5.3 Results

5.3.1 Identification of Aberrant Expression, Methylation and SCNA in Inflammation-related Genes

We first obtained the profiles of mRNA expression, DNA methylation and copy number changes of all inflammation-related genes utilizing the array-based techniques. Overall, we identified 260 transcripts covering 252 inflammation-related genes that exhibited substantially aberrant expression in HCCs. Among them, 121 transcripts (114 genes) were up-regulated while 139 (138 genes) were down-regulated. We then performed exploratory hierarchical clustering of the 260 transcripts and found that the expression profiles in tumors and adjacent non-tumor tissues created unequivocally separate clusters (Figure 5.1a, left). We found 71 probes residing in 39 inflammation-related gene loci were hypermethylated in HCCs, 391 probes in 85 genes hypomethylated. In addition, there were two differentially methylated probes in the FYN gene with one hypermethylated and the other hypomethylated. Hierarchical clustering resulted in overall clear distinction between tumor and adjacent non-tumor tissues, except for 3 samples (Figure 5.1a, middle). As to copy number profiles, we identified 131 inflammation-related genes (consistent evidence from 512 probes) showing significant copy number gain and 141 genes (consistent evidence from 617 probes) showing significant copy number loss in HCC tumor tissues. In addition, 117 probes showed significant copy number difference between tumor and non-tumor tissues, exhibiting effects in both directions across 15 genes. Hierarchical clustering based on copy number resulted in clear distinction between tumor and non-tumor tissues (Figure 5.1a, right).

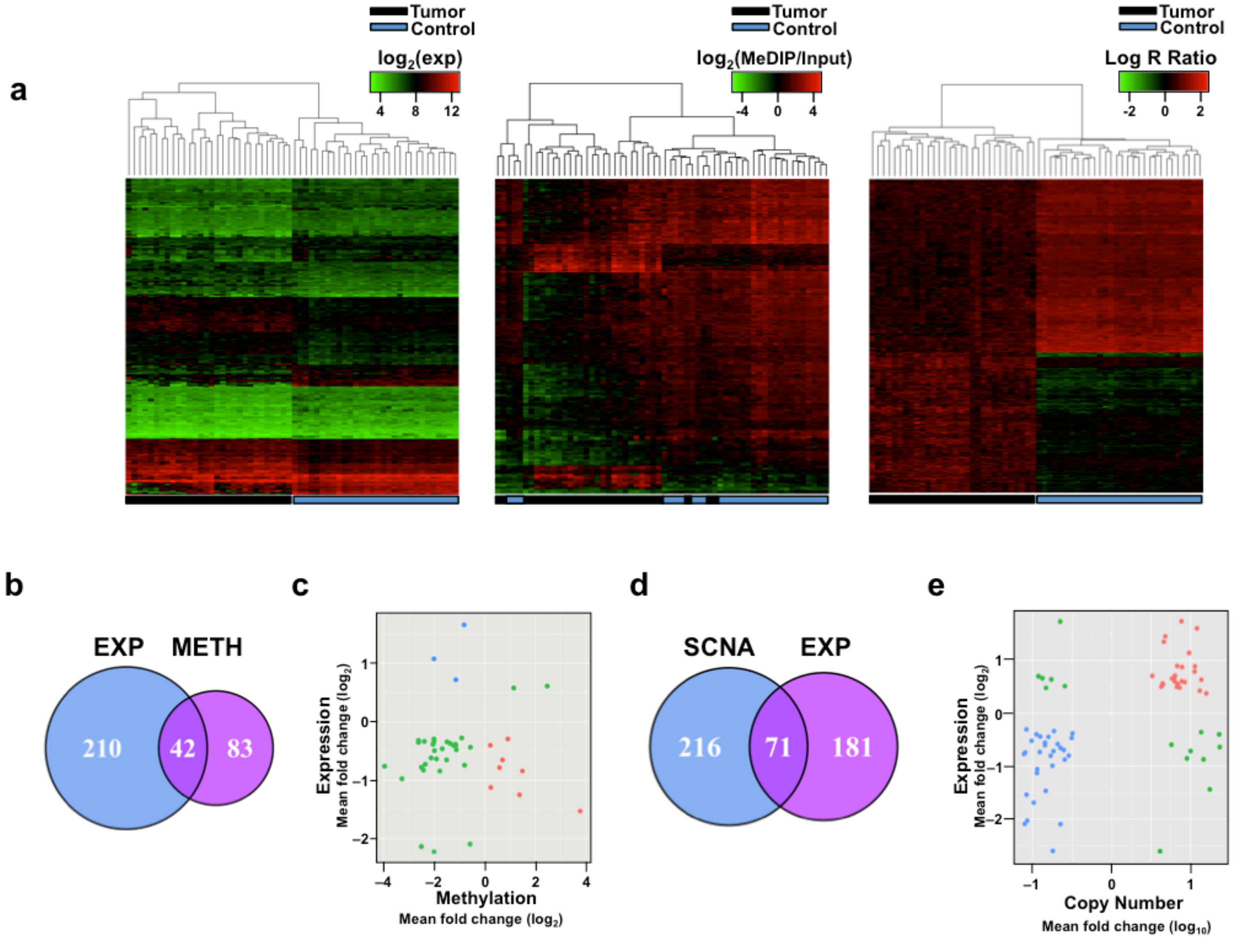


Figure 5.1 Identification of inflammation-related genes exhibiting coordinative changes between mRNA expression and DNA methylation or SCNA in HBV-related HCCs. (a) Hierarchical clustering with 260 significant transcripts corresponding 252 genes (left), 464 significant methylation probes covering 125 genes (middle), and 1,246 significant SNP/CN probes covering 287 genes (right) across HCCs (Tumor) and paired non-tumor tissues (Control). (b) Venn diagram showing 42 overlapping genes identified by the analysis of both mRNA expression and DNA methylation. EXP, genes with aberrant expression; METH, genes with aberrant DNA methylation. (c) Starburst plot of 42 overlapping genes identified by the analysis of both gene expression and DNA methylation. Red, blue and green dots indicate the genes hypermethylated and down-regulated, the genes hypomethylated and up-regulated and the genes having their expression not associated with DNA methylation, respectively. (d) Venn diagram showing 71 overlapping genes identified by the analysis of both mRNA expression and SCNA. EXP, genes with aberrant expression; SCNA, genes with copy number aberration. (e) Starburst plot of 71 overlapping genes identified by the analysis of both mRNA expression and SCNA. Red, blue and green dots indicate the genes with copy number gain and up-regulated expression, the genes with copy number deletion and down-regulated expression and the genes having inverse relationship between SCNA and mRNA expression, respectively.

5.3.2 Contribution of Methylation and SCNA to Aberrant Expression of Inflammation-related Genes

To investigate the contributions of DNA methylation and SCNA to the aberrant expression of inflammation-related genes, we integrated the expression profiles and DNA methylation profiles or copy number profiles obtained from tumors and non-tumor tissues. We found that 42 genes with aberrant expression had aberrant DNA methylation in HCCs compared with paired non-tumor tissues. Among them, only 8 genes with substantial down-regulation had DNA hypermethylation and 3 genes with substantial over-expression had DNA hypomethylation (Figure 5.1b and Figure 5.1c), indicating a possible minor role (< 5% of total) of DNA methylation in the expression regulation of these inflammation-related genes. We observed essential segregation of the expression levels and DNA methylation values in tumor and non-tumor tissues (Figure 5.1c). For the correlation between gene expression and SCNA, we found 56 genes with aberrant expression had concomitant SCNA in HCCs. Of these 56 genes, 26 with over-expression had copy number gain, while 30 with down-regulation had DNA deletion; SCNA can explain only one-fifth of aberrant expression of inflammation-related genes in HCCs (Figure 5.1d and Figure 5.1e). The essential segregation of the gene expression levels and DNA copy number values is shown in Figure 5.1e. We found a negative correlation between the expression levels and DNA methylation for 11 genes (Figure 5.2a) and a positive correlation between the expression levels and copy numbers for 56 genes (Figure 5.2b). Interestingly, not only aberrant DNA methylation but also SCNA contribute to the aberrant expression of BCL2, ESR1, FYN, PRKCB, and PTPN13 in HCCs.

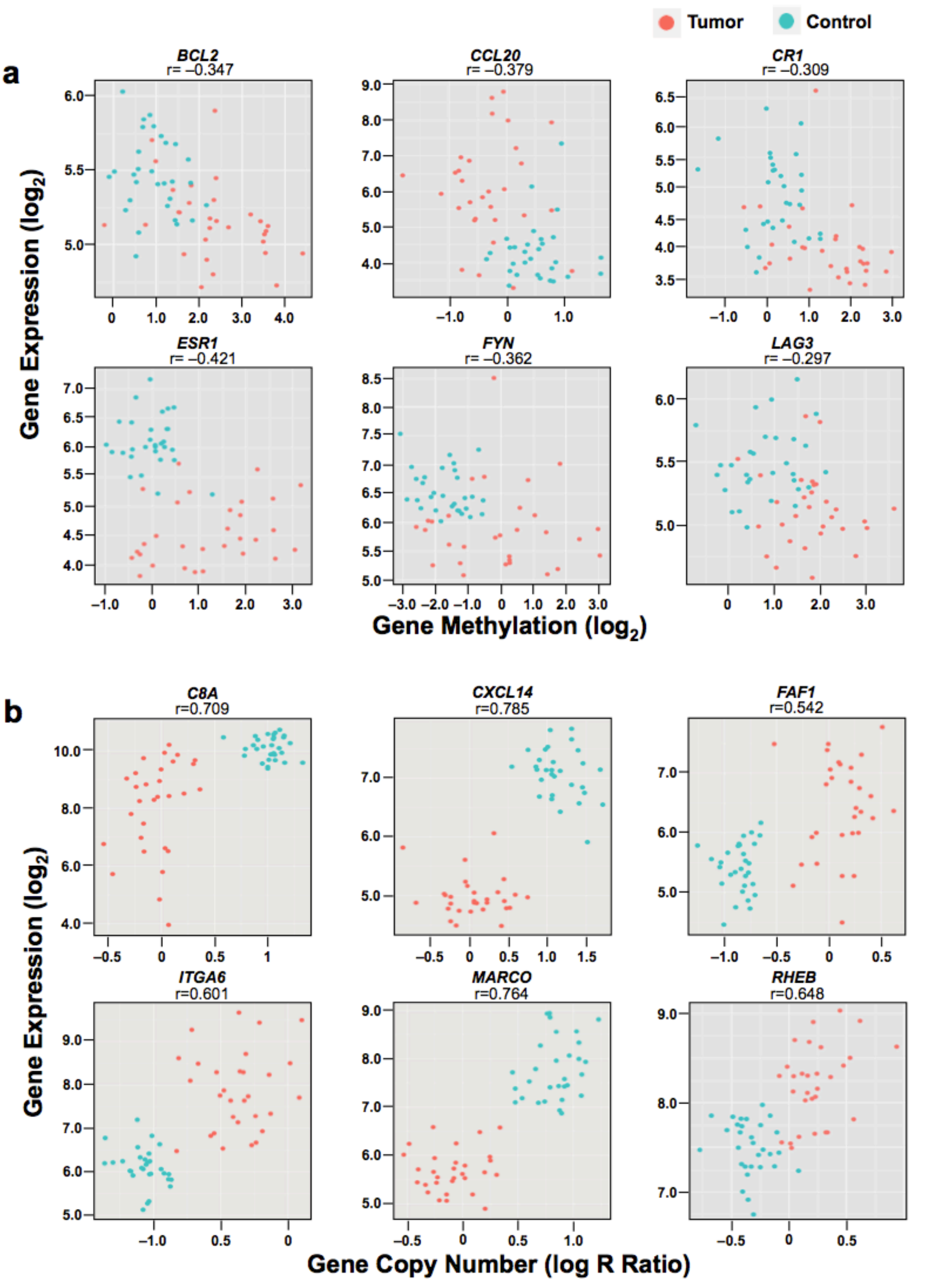


Figure 5.2 Relationship between mRNA expression and DNA methylation (a) or SCNA (b) of the randomly selected inflammation-related genes. Tumor, HCCs; Control, paired no-tumor tissues.

5.3.3 Validation of Methylation and SCNA Relevant to Aberrant Inflammation-related Gene Expression

There are currently no published integrative analyses on the relationship between aberrant expression of inflammation-related genes and DNA methylation or SCNA in HCCs. To fill in this gap, we investigated the overlapping genes with aberrant expression and aberrant DNA methylation or SCNA in our dataset and the public datasets. Of the 252 aberrantly expressed genes found in our study, 200 (79.4%) and 204 (81.0%) were verified in the GSE14520 and GSE25097 dataset, which totaled up to 237 (94.1%) in at least one dataset (Figure 5.3a). As to the 125 aberrantly methylated genes identified in our study, 110 (88.0%) and 47 (37.6%) were also found to be aberrantly methylated in the GSE54503 and GSE37988 dataset, respectively, totaling up to 88.8% in at least one dataset (Figure 5.3b). We compared SCNA identified in our study with that in the GSE38323 dataset, and found that among the 287 genes with SCNA, 68 (23.7%) were also reported in the public dataset. In addition, as to the 56 genes with aberrant expression and concomitant SCNA identified in our study, 17 (30.4%) were found in GSE28127.

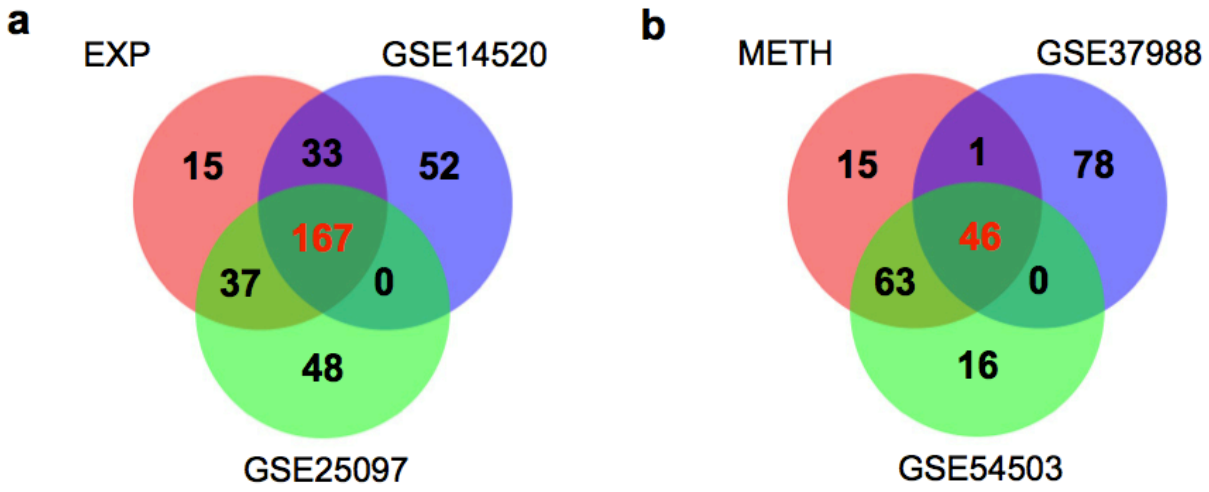


Figure 5.3 Venn diagram of the overlapping inflammation-related genes having significant mRNA expression (a), or DNA methylation (b) in HBV-related HCCs in our study with public datasets. EXP, genes with aberrant expression; METH, genes with aberrant methylation.

We next randomly selected 8 genes aberrantly expressed in HCCs that had aberrant DNA methylation (CR1, ESR1, PTPN13, and SOCS2) or SCNA (C8A, CXCL14, ITGA6 and MARCO) to validate the array-based results in an independent sample set consisting 47 HCCs and paired non-tumor specimens. The array-based analysis in our discovery stage showed that for PTPN13 and SOCS2, the methylation sites are located in the CpG islands but for CR1 and ESR1 the methylation sites are located in the intronic and promoter regions. MS-HRM was used to measure methylation at these 4 genes in the validation sample set. We found substantial changes of both DNA methylation and mRNA expression of CR1, ESR1, PTPN13, and SOCS2 in HCCs compared with paired non-tumor tissues (Figure 5.4a), which is consistent with the results obtained by analyzing the array data in our discovery sample. Similarly, we were able to validate the association between SCNA and gene expression of C8A, CXCL14, ITGA6 and MARCO (Figure 5.4b). These validation experiments indicate that the results produced by array-based analysis are reliable.

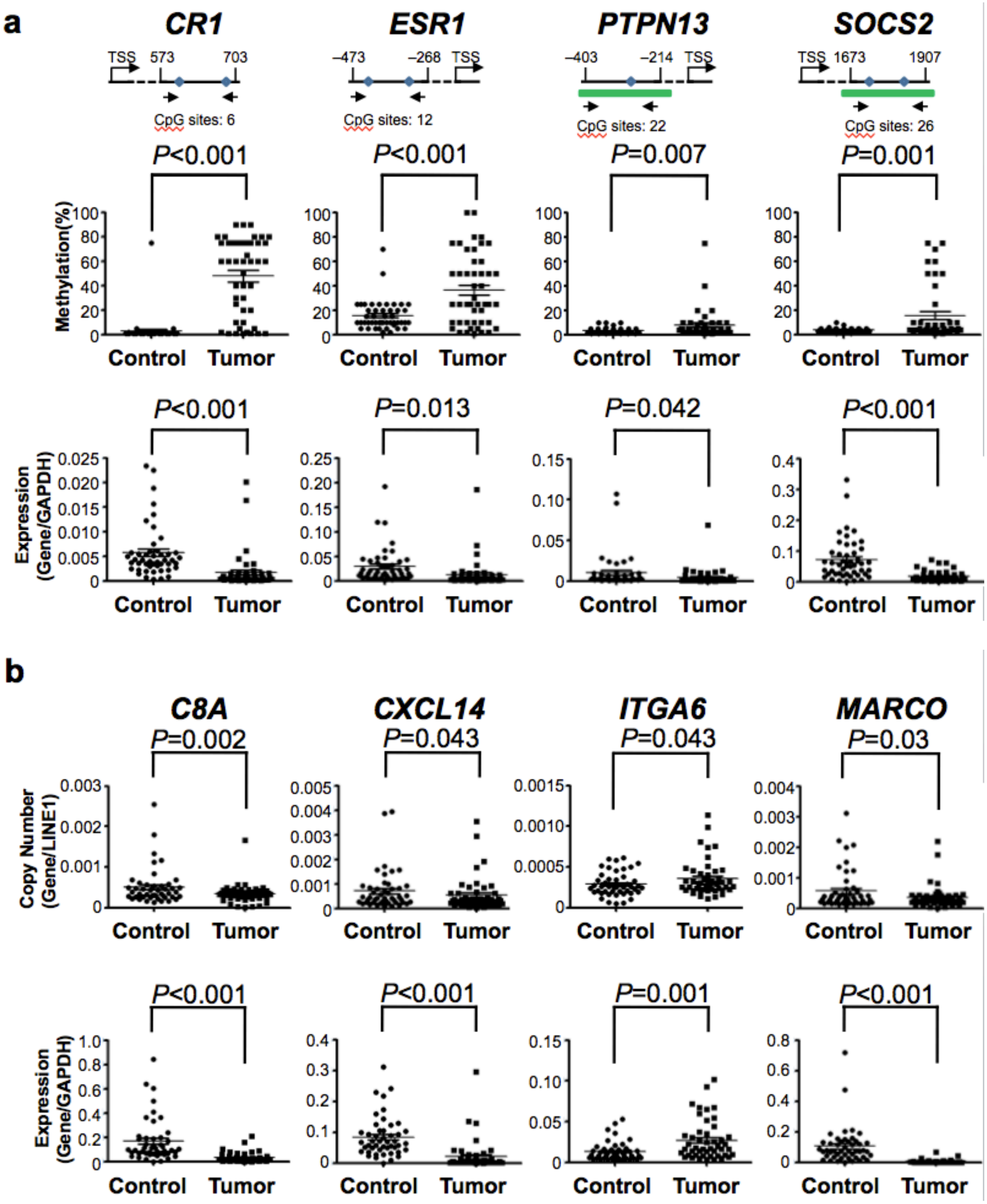


Figure 5.4 Validation of mRNA expression, DNA methylation and copy number aberration of the randomly selected inflammation-related genes in 47 samples. (a) Validations of mRNA expression and DNA methylation of CR1, ESR1, PTPN13 and SOCS2. TSS, transcription start site; Tumor, HCCs; Control, paired non-tumor tissues. Arrows represent the validation primers.

Blue dots represent the significant methylation probes designed in array. Green bars represent the CpG island. (b) Validation of gene expression and copy number aberration of C8A, CXCL14, ITGA6 and MARCO.

5.3.4 Functional Network Construction

We used MetaCore database and software to investigate the possible functional networks of inflammation-related genes suffering from aberrant DNA methylation (11 genes, Table 5.2) or SCNA (56 genes) in HCCs. For the genes suffering from aberrant methylation, the top functional network (gScore = 76.97) involved in the following GO processes: response to alcohol, response to organic cyclic compound, immune response-regulating cell surface receptor signaling pathway, response to steroid hormone stimulus and immune response-regulating signaling pathway, using PKC-β, Bcl-2, FAP-1, ESR and PKC as key nodes (Figure 5.5a). For the genes suffering from SCNA, the top functional network (gScore = 271.81) involved in the GO processes of immune response-regulating signaling pathway, Fc receptor signaling pathway, innate immune response, regulation of immune system process and regulation of immune response, using FLT3, GRB2, TCF8, β-catenin, and Fyn as key nodes (Figure 5.5b). Noticeably, the functional network of the genes suffering from aberrant methylation is distinct from that of the genes suffering from SCNA, suggesting that methylation and SCNA contribute to different inflammation processes which may have the different regulatory mechanism.

Table 5.2 Inflammation-related genes with DNA methylation changes associated with inverse expression changes in HCC

Symbol	HUGO Gene Name*	Location	Correlation (r)†	Methylation‡	Expression Changes‡
BCL2	B-cell CLL/lymphoma 2	18q21.3	–0.347	Hypermethylated	Down-regulated
CCL20	chemokine (C-C motif) ligand 20	2q36.3	–0.379	Hypomethylated	Up-regulated
CR1	complement component (3b/4b) receptor 1 (Knops blood group)	1q32	–0.309	Hypermethylated	Down-regulated
ESR1	estrogen receptor 1	6q24-q27	–0.421	Hypermethylated	Down-regulated
FYN	FYN oncogene related to SRC, FGR, YES	6q21	–0.362	Hypermethylated	Down-regulated
LAG3	lymphocyte-activation gene 3	12p13.3	–0.297	Hypermethylated	Down-regulated
NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog	1p13.2	–0.542	Hypomethylated	Up-regulated
PRKCB	protein kinase C, beta	16p12	–0.356	Hypermethylated	Down-regulated
PTPN13	protein tyrosine phosphatase, non-receptor type 13 (APO-1/CD95 (Fas)-associated phosphatase)	4q21.3	–0.556	Hypermethylated	Down-regulated
SOCS2	suppressor of cytokine signaling 2	12q	–0.408	Hypermethylated	Down-regulated
SPTAN1	spectrin, alpha, non-erythrocytic 1	9q34.11	–0.400	Hypomethylated	Up-regulated

* From HGNC database, http://www.genenames.org/.

†Correlation of DNA methylation and expression.

‡Results in this study.

Abbreviations: N/A, Not available.

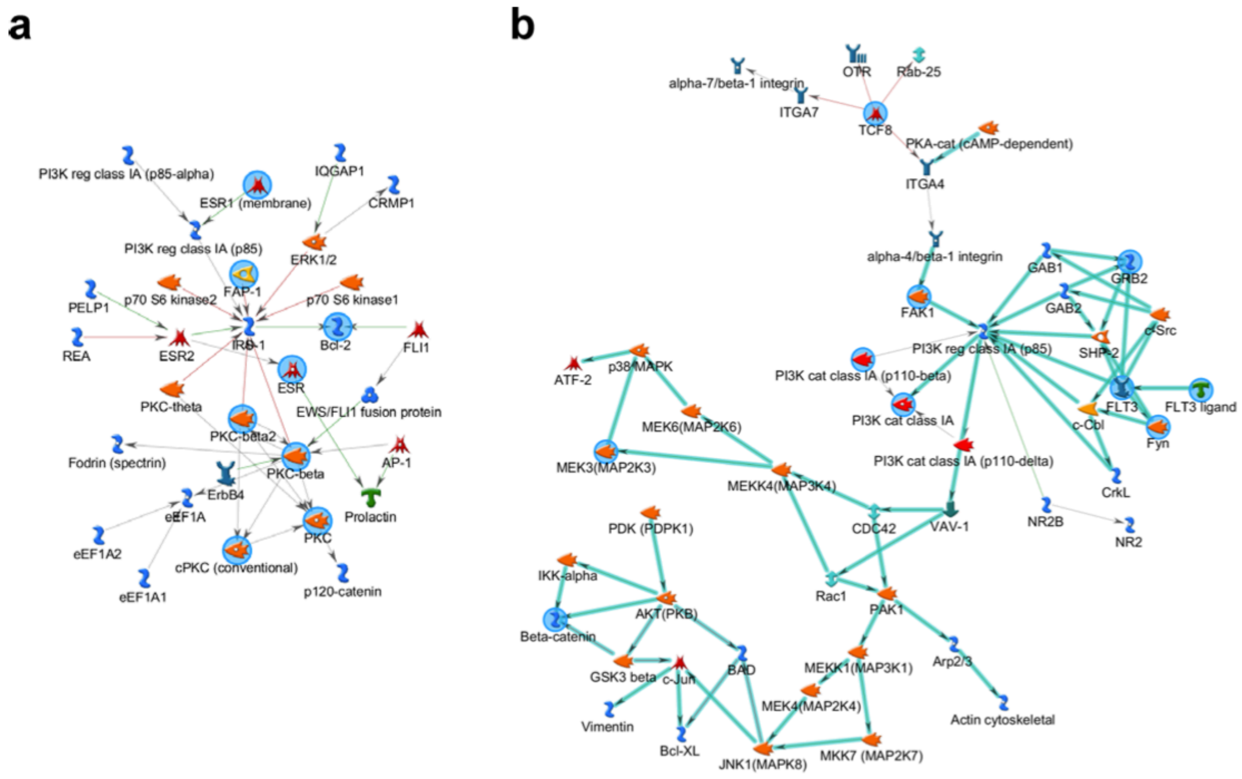


Figure 5.5 The functional network of top inflammation-related genes created by integrative analysis of mRNA expression associated with aberrant DNA methylation (a) and SCNA (b) in HBV-related HCCs. Blue circles represent the key nodes of the network.

5.4 Discussion

It is well known that chronic inflammation induced by HBV plays a pivotal role in the development of HCC [Grivennikov, et al. 2010; Rogers and Fox 2004]. Many efforts have been made to investigate the aberrant expression and the regulatory mechanisms of inflammation-related genes in HBV-related HCC, but most have focused on a single or few genes. In this study, we performed a two-stage analysis to systematically investigate the contributions of aberrant DNA methylation and SCNA to the aberrant expression of genes in the whole inflammation system.

We identified 252 differentially expressed, 125 aberrantly methylated and 287 copy number changed inflammation-related genes in HBV-related HCC, which were validated in several published datasets. Except for the GSE54503 dataset, all the other five referred datasets were obtained from Chinese patients with HCC, which may efficiently reduce ethnic bias for comparative analysis. Most of the differentially expressed genes (94.1%) and aberrantly methylated genes (88.8%) were found in at least one referred dataset, indicating consistent results among these studies and ours. However, only 23.7% of copy number-changed genes identified in our study were also identified in the published report of GSE38323, probably due to different array platforms and methods used to detect SCNA across studies.

The most significant result in the current study is that we integrated the expression profile with DNA methylation and SCNA profiles to investigate to what extent the aberrant expression of these inflammation genes is attributable to methylation or SCNA. We found that among the 252 aberrantly expressed inflammation genes in HCC, only 11 genes whose aberrant expression can be explained by change in methylation; and only 56 genes whose aberrant expression can be explained by SCNA. It has been well known that aberrant DNA methylation is one of the most crucial hallmarks of carcinogenesis. However, our results revealed that aberrant DNA methylation might play a minor role (<5%) in the transcriptional regulation of inflammation-related genes in HBV-related HCCs. Previous studies have shown that aberrant DNA methylation of inflammation-related genes comes into existence in liver cirrhosis and some are maintained in HCCs [Ammerpohl, et al. 2012; Komatsu, et al. 2012]. These findings are in agreement with the notion that both aberrant DNA methylation and inflammatory response occur as early events in HCC carcinogenesis. It has been reported that structure aberrations are persistently accumulated from early to late stage of HCCs [Trevisani, et al. 2008]. Our results indicated that SCNA might explain one-fifth aberrant expression of inflammation genes in HBV-related HCCs. Taken together, aberrant DNA methylation and SCNAs in HBV-related HCCs contributed to less than 30% aberrant expression of inflammation-related genes. Furthermore, analysis of functional networks of aberrantly expressed genes showed that aberrant DNA methylation and SCNA brought about different inflammatory response pathways. All these findings suggested that the regulatory system of inflammation-related genes is complicated and meticulous. Other molecular mechanisms such as transcription factors, non-coding RNAs and mutations might be more relevant in terms of aberrant expression of inflammation-related genes in HCCs, which warrants future investigation.

Four genes (CR1, ESR1, PTPN13, and SOCS2) that showed aberrant expression associated with DNA methylation were randomly selected to verify in an additional sample set and proved to be consistent with array-based results. CR1, a negative regulator of the complement cascade, has been reported to be hypermethylated in acute lymphoblastic leukemia [Nordlund, et al. 2012]. ESR1, a ligand-activated transcription factor, has been shown to be hypermethylated in the promoter in 83.3% of HCC samples [Hishida, et al. 2013]. PTPN13 is a protein tyrosine phosphatase gene that is frequently down-regulated and hypermethylated in HCC cell lines [Ying, et al. 2006]. SOCS2, encoding the suppressor of cytokine signaling protein, is frequently hypermethylated in primary ovarian cancer [Sutherland, et al. 2004]. Similarly, the association between SCNA and aberrant expression of C8A, CXCL14, ITGA6 and MARCO was also verified in the validation sample set. The CXCL14 gene, whose copy number was significantly lower in HCCs compared with paired non-tumor tissues, encodes a chemokine that has been shown to play a pivotal role as a tumor suppressor in HCC [Wang, et al. 2013c]. However, the roles that C8A, ITGA6 and MARCO genes play in the development of HCC remain unclear and further studies on these genes are under way.

We acknowledge some limitations of this study. Although tissue heterogeneity of the liver is much less than other tissues such as the lung and the breast, which provides good opportunity for analysis of gene expression in cancer tissue, the stromal content in clinical samples might also have potential to affect mRNA analysis of HCC. Besides, this study analyzed only the mRNA levels and it would be profited to analyze the protein level of some genes to confirm their mRNA levels.

In summary, our two-stage comprehensive study achieved at least two progresses: identification of HCC-specific DNA methylation, SCNA and mRNA expression profiles of inflammation-related genes and elucidation of the contribution of DNA methylation and SCNA to the aberrant expression of inflammation-related genes in HBV-related HCCs. These results partially answered the long-standing question, i.e., what mechanism contributes to the aberrant expression of inflammation-related genes that play important roles in the development of HBV-related HCCs.

CHAPTER 6 CONCLUDING REMARKS

This document presents several novel methods for exploring regulatory mechanisms of gene expression. While each method is intended for a specific study design and involves a variety of statistical and computational tools, the central goal remains the same: to identify true signals from the noisy omics data. We have demonstrated that various statistical methodology can be used to make better use of currently available data. For Hi-C data, we proposed a HMRF based Bayesian method and found that borrowing information from neighboring loci pairs via explicitly modeling the dependency of peak pattern in 2D space can improve power to detect long-range chromosomal interactions, particularly with high-resolution data. For DNA methylation data, we proposed a penalized functional regression model and demonstrated that by incorporating functional predictors, our model can utilize information from non-local probes to substantially improve the imputation quality and the statistical power to identify trait-associated methylation loci in downstream EWAS. For the integrative analysis of HCC, we demonstrated that multi-omics data can be aggregated to reveal regulatory mechanisms of aberrant expression of inflammation-related genes in HBV-related HCC.

In general, our methods have demonstrated key advantages in analyzing various omics datasets. With the continuing drop in costs of high-throughput experiments, we expect they will lead to rapid advances in understanding the diverse regulatory mechanisms of gene expression and the impact of their aberrations.

REFERENCES

Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, Pochanard P, Mozes E, Garraway LA, Pe'er D. 2010. An Integrated Approach to Uncover Drivers of Cancer. Cell 143(6):1005-1017.

Albertson DG, Collins C, McCormick F, Gray JW. 2003. Chromosome aberrations in solid tumors. Nat Genet 34(4):369-76.

Alitalo K, Schwab M, Lin CC, Varmus HE, Bishop JM. 1983. Homogeneously staining chromosomal regions contain amplified copies of an abundantly expressed cellular oncogene (c-myc) in malignant neuroendocrine cells from a human colon carcinoma. Proc Natl Acad Sci U S A 80(6):1707-11.

Ammerpohl O, Pratschke J, Schafmayer C, Haake A, Faber W, von Kampen O, Brosch M, Sipos B, von Schonfels W, Balschun K and others. 2012. Distinct DNA methylation patterns in cirrhotic liver and hepatocellular carcinoma. Int J Cancer 130(6):1319-28.

Ay F, Bailey TL, Noble WS. 2014. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Research 24:999-1011.

Barfield RT, Almli LM, Kilaru V, Smith AK, Mercer KB, Duncan R, Klengel T, Mehta D, Binder EB, Epstein MP and others. 2014. Accounting for population stratification in DNA methylation studies. Genet Epidemiol 38(3):231-41.

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D and others. 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483(7391):603-7.

Bau D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, Dekker J, Marti-Renom MA. 2011. The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. Nat Struct Mol Biol 18(1):107-14.

Baylin SB, Herman JG, Graff JR, Vertino PM, Issa JP. 1998. Alterations in DNA methylation: a fundamental aspect of neoplasia. Adv Cancer Res 72:141-96.

Bergman Y, Cedar H. 2013. DNA methylation dynamics in health and disease. Nature Structural & Molecular Biology 20(3):274-281.

Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu YP, Noushmehr H, Lange CPE, van Dijk CM, Tollenaar RAEM and others. 2012. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. Nature Genetics 44(1):40-U62.

Besag J, Green P, Higdon D, Mengersen K. 1995. Bayesian Computation and Stochastic-Systems. Statistical Science 10(1):3-41.

Bestor TH, Tycko B. 1996. Creation of genomic methylation patterns. Nat Genet 12(4):363-7.

Bhasin M, Zhang H, Reinherz EL, Reche PA. 2005. Prediction of methylated CpGs in DNA sequences using a support vector machine. FEBS Lett 579(20):4302-8.

Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL and others. 2011. High density DNA methylation array with single CpG site resolution. Genomics 98(4):288-295.

Bird A. 2002. DNA methylation patterns and epigenetic memory. Genes & Development 16(1):6-21.

Bo TH, Dysvik J, Jonassen I. 2004. LSimpute: accurate estimation of missing values in microarray data with least squares methods. Nucleic Acids Research 32(3).

Bock C, Paulsen M, Tierling S, Mikeska T, Lengauer T, Walter J. 2006. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. PLoS Genet 2(3):e26.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19(2):185-93.

Brandeis M, Frank D, Keshet I, Siegfried Z, Mendelsohn M, Nemes A, Temper V, Razin A, Cedar H. 1994. Sp1 elements protect a CpG island from de novo methylation. Nature 371(6496):435-8.

Cancer Genome Atlas Research N. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455(7216):1061-8.

Cancer Genome Atlas Research N. 2013. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499(7456):43-9.

Cavenee WK, Dryja TP, Phillips RA, Benedict WF, Godbout R, Gallie BL, Murphree AL, Strong LC, White RL. 1983. Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. Nature 305(5937):779-84.

Cedar H, Bergman Y. 2012. Programming of DNA methylation patterns. Annu Rev Biochem 81:97-117.

Chen Y, Ning Y, Hong C, Wang S. 2014. Semiparametric tests for identifying differentially methylated loci with case-control designs using Illumina arrays. Genet Epidemiol 38(1):42-50.

Choi H, Qin ZS, Ghosh D. 2010. A double-layered mixture model for the joint analysis of DNA copy number and gene expression data. J Comput Biol 17(2):121-37.

Choy MK, Movassagh M, Goh HG, Bennett MR, Down TA, Foo RS. 2010. Genome-wide conserved consensus transcription factor binding motifs are hyper-methylated. BMC Genomics 11:519.

Cleveland WS. 1979. Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association 74(368):829-836.

Consortium EP. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57-74.

Das R, Dimitrova N, Xuan Z, Rollins RA, Haghighi F, Edwards JR, Ju J, Bestor TH, Zhang MQ. 2006. Computational prediction of methylation status in human genomic sequences. Proc Natl Acad Sci U S A 103(28):10713-6.

de Wit E, de Laat W. 2012. A decade of 3C technologies: insights into nuclear organization. Genes Dev 26(1):11-24.

Deaton AM, Bird A. 2011. CpG islands and the regulation of transcription. Genes Dev 25(10):1010-22.

Dekker J, Marti-Renom MA, Mirny LA. 2013. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet 14(6):390-403.

Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. Science 295(5558):1306-1311.

Dickson J, Gowher H, Strogantsev R, Gaszner M, Hair A, Felsenfeld G, West AG. 2010. VEZF1 elements mediate protection from DNA methylation. PLoS Genet 6(1):e1000804.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485(7398):376-80.

Donner Y, Feng T, Benoist C, Koller D. 2012. Imputing gene expression from selectively reduced probe sets. Nat Methods 9(11):1120-5.

Duan Q, Liu EY, Auer PL, Zhang G, Lange EM, Jun G, Bizon C, Jiao S, Buyske S, Franceschini N and others. 2013. Imputation of coding variants in African Americans: better performance using data from the exome sequencing project. Bioinformatics 29(21):2744-9.

Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS. 2010. A three-dimensional model of the yeast genome. Nature 465(7296):363-367.

Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA and others. 2006. DNA methylation profiling of human chromosomes 6, 20 and 22. Nature Genetics 38(12):1378-1385.

El-Serag HB, Rudolph KL. 2007. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. Gastroenterology 132(7):2557-76.

Ernst J, Kellis M. 2015. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat Biotechnol 33(4):364-76.

Fan R, Wang Y, Boehnke M, Chen W, Li Y, Ren H, Lobach I, Xiong M. 2015a. Gene Level Meta-Analysis of Quantitative Traits by Functional Linear Models. Genetics 200(4):1089-104.

Fan R, Wang Y, Chiu CY, Chen W, Ren H, Li Y, Boehnke M, Amos CI, Moore JH, Xiong M. 2015b. Meta-analysis of Complex Diseases at Gene Level by Generalized Functional Linear Models. Genetics.

Fan SC, Zhang MQ, Zhang XG. 2008. Histone methylation marks play important roles in predicting the methylation status of CpG islands. Biochemical and Biophysical Research Communications 374(3):559-564.

Fang F, Fan SC, Zhang XG, Zhang MQ. 2006. Predicting methylation status of CpG islands in the human brain. Bioinformatics 22(18):2204-2209.

Fattovich G, Stroffolini T, Zagni I, Donato F. 2004. Hepatocellular carcinoma in cirrhosis: incidence and risk factors. Gastroenterology 127(5 Suppl 1):S35-50.

François O, Ancelet S, Guillot G. 2006. Bayesian clustering using hidden Markov random fields in spatial population genetics. Genetics 174(2):805-816.

Freudenberg J, Lee HS, Han BG, Shin HD, Kang YM, Sung YK, Shim SC, Choi CB, Lee AT, Gregersen PK and others. 2011. Genome-Wide Association Study of Rheumatoid Arthritis in Koreans. Arthritis and Rheumatism 63(4):884-893.

Gaszner M, Felsenfeld G. 2006. Insulators: exploiting transcriptional and epigenetic mechanisms. Nat Rev Genet 7(9):703-13.

Gelman A. 2004. Bayesian data analysis. Boca Raton, Fla.: Chapman & Hall/CRC.

Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56-65.

Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R and others. 2012. Architecture of the human regulatory network derived from ENCODE data. Nature 489(7414):91-100.

Gervin K, Hammero M, Akselsen HE, Moe R, Nygard H, Brandt I, Gjessing HK, Harris JR, Undlien DE, Lyle R. 2011. Extensive variation and low heritability of DNA methylation identified in a twin study. Genome Res 21(11):1813-21.

Getz G, Gabriel SB, Cibulskis K, Lander E, Sivachenko A, Sougnez C, Lawrence M, Kandoth C, Dooling D, Fulton R and others. 2013. Integrated genomic characterization of endometrial carcinoma. Nature 497(7447):67-73.

Gibcus JH, Dekker J. 2013. The hierarchy of the 3D genome. Mol Cell 49(5):773-82.

Goldsmith J, Bobb J, Crainiceanu CM, Caffo B, Reich D. 2011. Penalized Functional Regression. Journal of Computational and Graphical Statistics 20(4):830-851.

Gonzalo S. 2010. Epigenetic alterations in aging. Journal of Applied Physiology 109(2):586-597.

Gorre ME, Mohammed M, Ellwood K, Hsu N, Paquette R, Rao PN, Sawyers CL. 2001. Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. Science 293(5531):876-80.

Grivennikov SI, Greten FR, Karin M. 2010. Immunity, inflammation, and cancer. Cell 140(6):883-99.

Hamers L, Hemeryck Y, Herweyers G, Janssen M, Keters H, Rousseau R, Vanhoutte A. 1989. Similarity Measures in Scientometric Research - the Jaccard Index Versus Salton Cosine Formula. Information Processing & Management 25(3):315-318.

Hanahan D, Weinberg RA. 2011. Hallmarks of cancer: the next generation. Cell 144(5):646-74.

Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D and others. 2011. Increased methylation variation in epigenetic domains across cancer types. Nat Genet 43(8):768-75.

Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong CB, Downey SL, Johnson BE, Fouse SD, Delaney A, Zhao YJ and others. 2010. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nature Biotechnology 28(10):1097-U194.

Hawkins RD, Hon GC, Lee LK, Ngo Q, Lister R, Pelizzola M, Edsall LE, Kuan S, Luu Y, Klugman S and others. 2010. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell 6(5):479-91.

Heyn H, Esteller M. 2012. DNA methylation profiling in the clinic: applications and challenges. Nature Reviews Genetics 13(10):679-692.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America 106(23):9362-9367.

Hishida M, Nomoto S, Inokawa Y, Hayashi M, Kanda M, Okamura Y, Nishikawa Y, Tanaka C, Kobayashi D, Yamada S and others. 2013. Estrogen receptor 1 gene as a tumor suppressor gene in hepatocellular carcinoma detected by triple-combination array analysis. International Journal of Oncology 43(1):88-94.

Horvath S. 2013. DNA methylation age of human tissues and cell types. Genome Biol 14(10):R115.

Hou CH, Li L, Qin ZHS, Corces VG. 2012. Gene Density, Transcription, and Insulators Contribute to the Partition of the Drosophila Genome into Physical Domains. Molecular Cell 48(3):471-484.

Hu M, Deng K, Qin ZH, Dixon J, Selvaraj S, Fang J, Ren B, Liu JS. 2013. Bayesian Inference of Spatial Organizations of Chromosomes. Plos Computational Biology 9(1).

Huang KC, Sun W, Wu Y, Chen M, Mohlke KL, Lange LA, Li Y. 2014. Association studies with imputed variants using expectation-maximization likelihood-ratio tests. PLoS One 9(11):e110679.

Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature Methods 9(10):999-+.

Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M and others. 2009. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet 41(2):178-86.

Jewett EM, Zawistowski M, Rosenberg NA, Zollner S. 2012. A coalescent model for genotype imputation. Genetics 191(4):1239-55.

Jhunjhunwala S, van Zelm MC, Peak MM, Cutchin S, Riblet R, van Dongen JJM, Grosveld FG, Knoch TA, Murre C. 2008. The 3D structure of the immunoglobulin heavy-chain locus: Implications for long-range genomic interactions. Cell 133(2):265-279.

Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen CA, Schmitt AD, Espinoza CA, Ren B. 2013. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature 503(7475):290-4.

Johnson WE, Li C, Rabinovic A. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8(1):118-127.

Jones PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13(7):484-92.

Jones PA, Laird PW. 1999. Cancer epigenetics comes of age. Nat Genet 21(2):163-7.

Kim H, Golub GH, Park H. 2005. Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics 21(2):187-198.

Kindermann R, Snell JL, American Mathematical Society. 1980. Markov random fields and their applications. Providence, R.I.: American Mathematical Society.

Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, Fulton LL, Dooling DJ, Ding L, Mardis ER and others. 2012. Comprehensive molecular portraits of human breast tumours. Nature 490(7418):61-70.

Koivisto P, Kononen J, Palmberg C, Tammela T, Hyytinen E, Isola J, Trapman J, Cleutjens K, Noordzij A, Visakorpi T and others. 1997. Androgen receptor gene amplification: a possible molecular mechanism for androgen deprivation therapy failure in prostate cancer. Cancer Res 57(2):314-9.

Komatsu Y, Waku T, Iwasaki N, Ono W, Yamaguchi C, Yanagisawa J. 2012. Global analysis of DNA methylation in early-stage liver fibrosis. BMC Med Genomics 5:5.

Kumar V, Wijmenga C, Withoff S. 2012. From genome-wide association studies to disease mechanisms: celiac disease as a model for autoimmune diseases. Seminars in Immunopathology 34(4):567-580.

Lamb JR, Zhang C, Xie T, Wang K, Zhang B, Hao K, Chudin E, Fraser HB, Millstein J, Ferguson M and others. 2011. Predictive genes in adjacent normal tissue are preferentially altered by sCNV during tumorigenesis in liver cancer and may rate limiting. PLoS One 6(7):e20090.

Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Sung KWK, Rigoutsos I, Loring J and others. 2010. Dynamic changes in the human methylome during differentiation. Genome Research 20(3):320-331.

Law JA, Jacobsen SE. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet 11(3):204-20.

Lee JS. 2013. Genomic profiling of liver cancer. Genomics Inform 11(4):180-5.

Ley TJ, Miller C, Ding L, Raphael BJ, Mungall AJ, Robertson AG, Hoadley K, Triche TJ, Laird PW, Baty JD and others. 2013. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. New England Journal of Medicine 368(22):2059-2074.

Li GL, Ruan XA, Auerbach RK, Sandhu KS, Zheng MZ, Wang P, Poh HM, Goh Y, Lim J, Zhang JY and others. 2012. Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. Cell 148(1-2):84-98.

Li HZ, Wei Z, Maris J. 2010. A hidden Markov random field model for genome-wide association studies. Biostatistics 11(1):139-150.

Li J, Yen C, Liaw D, Podsypanina K, Bose S, Wang SI, Puc J, Miliaresis C, Rodgers L, McCombie R and others. 1997. PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. Science 275(5308):1943-7.

Li Y, Willer C, Sanna S, Abecasis G. 2009. Genotype Imputation. Annual Review of Genomics and Human Genetics 10:387-406.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO and others. 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326(5950):289-293.

Liew AWC, Law NF, Yan H. 2011. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Briefings in Bioinformatics 12(5):498-513.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM and others. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462(7271):315-322.

Liu EY, Buyske S, Aragaki AK, Peters U, Boerwinkle E, Carlson C, Carty C, Crawford DC, Haessler J, Hindorff LA and others. 2012. Genotype Imputation of MetabochipSNPs Using a Study-Specific Reference Panel of ~4,000 Haplotypes in African Americans From the Women's Health Initiative. Genetic Epidemiology 36(2):107-117.

Loza MJ, McCall CE, Li L, Isaacs WB, Xu J, Chang BL. 2007. Assembly of inflammation-related genes for pathway-focused genetic analysis. PLoS One 2(10):e1035.

Ma B, Wilker EH, Willis-Owen SA, Byun HM, Wong KC, Motta V, Baccarelli AA, Schwartz J, Cookson WO, Khabbaz K and others. 2014. Predicting DNA methylation level across human tissues. Nucleic Acids Res 42(6):3515-28.

Macleod D, Charlton J, Mullins J, Bird AP. 1994. Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes Dev 8(19):2282-92.

Maniatis T, Reed R. 2002. An extensive network of coupling among gene expression machines. Nature 416(6880):499-506.

Marti-Renom MA, Mirny LA. 2011. Bridging the Resolution Gap in Structural Modeling of 3D Genome Organization. Plos Computational Biology 7(7).

Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y and others. 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 466(7303):253-7.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu HZ, Brody J and others. 2012. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science 337(6099):1190-1195.

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB and others. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454(7205):766-70.

Mo QX. 2012. A fully Bayesian hidden Ising model for ChIP-seq data analysis. Biostatistics 13(1):113-128.

Moore MJ. 2005. From birth to death: The complex lives of eukaryotic mRNAs. Science 309(5740):1514-1518.

Nagai MA, Yamamoto L, Salaorni S, Pacheco MM, Brentani MM, Barbosa EM, Brentani RR, Mazoyer S, Smith SA, Ponder BA and others. 1994. Detailed deletion mapping of chromosome segment 17q12-21 in sporadic breast tumours. Genes Chromosomes Cancer 11(1):58-62.

Natoli G, Ghisletti S, Barozzi I. 2011. The genomic landscapes of inflammation. Genes Dev 25(2):101-6.

Network CGAR. 2012. Comprehensive genomic characterization of squamous cell lung cancers The Cancer Genome Atlas Research Network (vol 489, pg 519, 2012). Nature 491(7423):288-288.

Niu L, Li GL, Lin SL. 2014. Statistical Models for Detecting Differential Chromatin Interactions Mediated by a Protein. Plos One 9(5).

Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J and others. 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485(7398):381-385.

Nordlund J, Milani L, Lundmark A, Lonnerholm G, Syvanen AC. 2012. DNA Methylation Analysis of Bone Marrow Cells at Diagnosis of Acute Lymphoblastic Leukemia and at Remission. Plos One 7(4).

Orlow I, Lacombe L, Hannon GJ, Serrano M, Pellicer I, Dalbagni G, Reuter VE, Zhang ZF, Beach D, Cordon-Cardo C. 1995. Deletion of the p16 and p15 genes in human bladder tumors. J Natl Cancer Inst 87(20):1524-9.

Orphanides G, Reinberg D. 2002. A unified theory of gene expression. Cell 108(4):439-451.

Ozen C, Yildiz G, Dagcan AT, Cevik D, Ors A, Keles U, Topel H, Ozturk M. 2013. Genetics and epigenetics of liver cancer. N Biotechnol 30(4):381-4.

Pennisi E. 2011. The Biology of Genomes. Disease risk links to gene regulation. Science 332(6033):1031.

Phillips JE, Corces VG. 2009. CTCF: master weaver of the genome. Cell 137(7):1194-211.

Pistis G, Porcu E, Vrieze SI, Sidore C, Steri M, Danjou F, Busonero F, Mulas A, Zoledziewska M, Maschio A and others. 2015. Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. Eur J Hum Genet 23(7):975-83.

Previti C, Harari O, Zwir I, del Val C. 2009. Profile analysis and prediction of tissue-specific CpG island methylation classes. Bmc Bioinformatics 10.

Qin ZHS, Yu JJ, Shen JC, Maher CA, Hu M, Kalyana-Sundaram S, Yu JD, Chinnaiyan AM. 2010. HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. Bmc Bioinformatics 11.

Rakyan VK, Down TA, Balding DJ, Beck S. 2011. Epigenome-wide association studies for common human diseases. Nature Reviews Genetics 12(8):529-541.

Rogers AB, Fox JG. 2004. Inflammation and Cancer. I. Rodent models of infectious gastrointestinal and liver cancer. Am J Physiol Gastrointest Liver Physiol 286(3):G361-6.

Russel D, Lasker K, Webb B, Velazquez-Muriel J, Tjioe E, Schneidman-Duhovny D, Peterson B, Sali A. 2012. Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. Plos Biology 10(1).

Sajan SA, Hawkins RD. 2012. Methods for Identifying Higher-Order Chromatin Structure. Annual Review of Genomics and Human Genetics, Vol 13 13:59-82.

Santarius T, Shipley J, Brewer D, Stratton MR, Cooper CS. 2010. A census of amplified and overexpressed human cancer genes. Nat Rev Cancer 10(1):59-64.

Sanyal A, Lajoie BR, Jain G, Dekker J. 2012. The long-range interaction landscape of gene promoters. Nature 489(7414):109-U127.

Scarano MI, Strazzullo M, Matarazzo MR, D'Esposito M. 2005. DNA methylation 40 years later: Its role in human health and disease. J Cell Physiol 204(1):21-35.

Schimke RT, Kaufman RJ, Alt FW, Kellems RF. 1978. Gene amplification and drug resistance in cultured murine cells. Science 202(4372):1051-5.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. Cell 148(3):458-472.

Sheffield NC, Thurman RE, Song LY, Safi A, Stamatoyannopoulos JA, Lenhard B, Crawford GE, Furey TS. 2013. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. Genome Research 23(5):777-788.

Shen L, Kondo Y, Guo Y, Zhang J, Zhang L, Ahmed S, Shu J, Chen X, Waterland RA, Issa JP. 2007. Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. PLoS Genet 3(10):2023-36.

Shen RL, Olshen AB, Ladanyi M. 2009. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics 25(22):2906-2912.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou MM, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S and others. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research 15(8):1034-1050.

Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W. 2006. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet 38(11):1348-54.

Slamon DJ, Godolphin W, Jones LA, Holt JA, Wong SG, Keith DE, Levin WJ, Stuart SG, Udove J, Ullrich A and others. 1989. Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. Science 244(4905):707-12.

Smallwood A, Ren B. 2013. Genome organization and long-range regulation of gene expression by enhancers. Current Opinion in Cell Biology 25(3):387-394.

Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gomez-Marin C, Aneas I, Credidio FL, Sobreira DR, Wasserman NF and others. 2014. Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature 507(7492):371-+.

Smith ZD, Meissner A. 2013. DNA methylation: roles in mammalian development. Nature Reviews Genetics 14(3):204-220.

Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, Lee BT, Learned K, Karolchik D, Hinrichs AS and others. 2016. The UCSC Genome Browser database: 2016 update. Nucleic Acids Res 44(D1):D717-25.

Spitz F, Furlong EE. 2012. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet 13(9):613-26.

Splinter E, de Laat W. 2011. The complex transcription regulatory landscape of our genome: control in three dimensions. EMBO J 30(21):4345-55.

Stingo FC, Vannucci M. 2011. Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. Bioinformatics 27(4):495-501.

Stratton MR, Campbell PJ, Futreal PA. 2009. The cancer genome. Nature 458(7239):719-24.

Sutherland KD, Lindeman GJ, Choong DYH, Wittlin S, Brentzell L, Phillips W, Campbell IG, Visvader JE. 2004. Differential hypermethylation of SOCS genes in ovarian and breast carcinomas. Oncogene 23(46):7726-7733.

Tanaka M, Katayama F, Kato H, Tanaka H, Wang J, Qiao YL, Inoue M. 2011. Hepatitis B and C virus infection and hepatocellular carcinoma in China: a review of epidemiology and control measures. J Epidemiol 21(6):401-16.

Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Gayther SA, Apostolidou S, Jones A, Lechner M, Beck S, Jacobs IJ and others. 2009. An epigenetic signature in peripheral blood predicts active ovarian cancer. PLoS One 4(12):e8274.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B and others. 2012. The accessible chromatin landscape of the human genome. Nature 489(7414):75-82.

Tost J. 2010. DNA methylation: an introduction to the biology and the disease-associated changes of a promising biomarker. Mol Biotechnol 44(1):71-81.

Trevisani F, Cantarini MC, Wands JR, Bernardi M. 2008. Recent advances in the natural history of hepatocellular carcinoma. Carcinogenesis 29(7):1299-305.

Trieu T, Cheng J. 2014. Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. Nucleic Acids Res.

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17(6):520-525.

Valenzuela L, Kamakaka RT. 2006. Chromatin insulators. Annu Rev Genet 40:107-38.

Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, Cross MK, Williams BA, Stamatoyannopoulos JA, Crawford GE and others. 2013. Dynamic DNA methylation across diverse human cell lines and tissues. Genome Research 23(3):555-567.

Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu JC, Haussler D, Stuart JM. 2010. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 26(12):i237-i245.

Wahl GM, Padgett RA, Stark GR. 1979. Gene amplification causes overproduction of the first three enzymes of UMP synthesis in N-(phosphonacetyl)-L-aspartate-resistant hamster cells. J Biol Chem 254(17):8679-89.

Wang J, Zhuang J, Iyer S, Lin XY, Greven MC, Kim BH, Moore J, Pierce BG, Dong X, Virgil D and others. 2013a. Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. Nucleic Acids Res 41(Database issue):D171-6.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M. 2007. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17(11):1665-74.

Wang K, Lim HY, Shi S, Lee J, Deng S, Xie T, Zhu Z, Wang Y, Pocalyko D, Yang WJ and others. 2013b. Genomic landscape of copy number aberrations enables the identification of oncogenic drivers in hepatocellular carcinoma. Hepatology 58(2):706-17.

Wang TL, Maierhofer C, Speicher MR, Lengauer C, Vogelstein B, Kinzler KW, Velculescu VE. 2002. Digital karyotyping. Proc Natl Acad Sci U S A 99(25):16156-61.

Wang WL, Huang PF, Zhang LF, Wei JF, Xie QS, Sun Q, Zhou XH, Xie HY, Zhou L, Zheng SS. 2013c. Antitumor efficacy of C-X-C motif chemokine ligand 14 in hepatocellular carcinoma in vitro and in vivo. Cancer Science 104(11):1523-1531.

Ward LD, Kellis M. 2012. Interpreting noncoding genetic variation in complex traits and human disease. Nature Biotechnology 30(11):1095-1106.

Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, Rebhan M, Schubeler D. 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet 39(4):457-66.

Wei P, Pan W. 2010. Network-based genomic discovery: application and comparison of Markov random-field models. Journal of the Royal Statistical Society Series C-Applied Statistics 59:105-125.

Witten DM, Tibshirani RJ. 2009. Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data. Statistical Applications in Genetics and Molecular Biology 8(1).

Wojdacz TK, Dobrovic A, Hansen LL. 2008. Methylation-sensitive high-resolution melting. Nat Protoc 3(12):1903-8.

Ying J, Li H, Cui Y, Wong AHY, Langford C, Tao Q. 2006. Epigenetic disruption of two proapoptotic genes MAPK10/JNK3 and PTPN13/FAP-1 in multiple lymphomas and carcinomas through hypermethylation of a common bidirectional promoter. Leukemia 20(6):1173-1175.

Zhang W, Liu Y, Sun N, Wang D, Boyd-Kirkup J, Dou XY, Han JDJ. 2013a. Integrating Genomic, Epigenomic, and Transcriptomic Features Reveals Modular Signatures Underlying Poor Prognosis in Ovarian Cancer. Cell Reports 4(3):542-553.

Zhang W, Spector TD, Deloukas P, Bell JT, Engelhardt BE. 2015. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. Genome Biol 16:14.

Zhang W, Zhu J, Schadt EE, Liu JS. 2010. A Bayesian Partition Method for Detecting Pleiotropic and Epistatic eQTL Modules. Plos Computational Biology 6(1).

Zhang Y, Wong CH, Birnbaum RY, Li G, Favaro R, Ngan CY, Lim J, Tai E, Poh HM, Wong E and others. 2013b. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature 504(7479):306-10.

Zhang Z. 2012. Genomic landscape of liver cancer. Nat Genet 44(10):1075-7.

Zheng H, Wu H, Li J, Jiang SW. 2013. CpGIMethPred: computational model for predicting methylation status of CpG islands in human genome. BMC Med Genomics 6 Suppl 1:S13.

Zhou X, Li ZC, Dai Z, Zou XY. 2012. Prediction of methylation CpGs and their methylation degrees in human DNA sequences. Computers in Biology and Medicine 42(4):408-413.

Ziller MJ, Gu HC, Muller F, Donaghey J, Tsai LTY, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE and others. 2013. Charting a dynamic DNA methylation landscape of the human genome. Nature 500(7463):477-481.

Zitvogel L, Kepp O, Galluzzi L, Kroemer G. 2012. Inflammasomes in carcinogenesis and anticancer immune responses. Nat Immunol 13(4):343-51.