4-12-2018 2:00 PM

Computational Modelling of Human Transcriptional Regulation by an Information Theory-based Approach

Ruipeng Lu
The University of Western Ontario

Supervisor
Rogan, Peter K.
The University of Western Ontario

Graduate Program in Computer Science
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Abstract

ChIP-seq experiments can identify the genome-wide binding site motifs of a transcription factor (TF) and determine its sequence specificity. Multiple algorithms were developed to derive TF binding site (TFBS) motifs from ChIP-seq data, including the entropy minimization-based Bipad that can derive both contiguous and bipartite motifs. Prior studies applying these algorithms to ChIP-seq data only analyzed a small number of top peaks with the highest signal strengths, biasing their resultant position weight matrices (PWMs) towards consensus-like, strong binding sites; nor did they derive bipartite motifs, disabling the accurate modelling of binding behavior of dimeric TFs.

This thesis presents a novel motif discovery pipeline by adding the recursive masking and thresholding functionalities to Bipad to improve detection of primary binding motifs. Analyzing 765 ENCODE ChIP-seq datasets with this pipeline generated contiguous and bipartite information theory-based PWMs (iPWMs) for 93 sequence-specific TFs, discovered 23 cofactor motifs for 127 TFs and revealed six high-confidence novel motifs. The accuracy of these iPWMs were determined via four independent validation methods, including detection of experimentally proven TFBSs, explanation of effects of characterized SNPs, comparison with previously published motifs and statistical analyses. Novel cofactor motifs supported previously unreported TF coregulatory interactions. This thesis further presents a unified framework to identify variants in hereditary breast and ovarian cancer (HBOC), successfully applying these iPWMs to prioritize TFBS variants in 20 complete genes of HBOC patients.

The spatial distribution and information composition of *cis*-regulatory modules (e.g. TFBS clusters) in promoters substantially determine gene expression patterns and TF target genes. Multiple algorithms were developed to detect TFBS clusters, including the information density-based clustering (IDBC) algorithm that simultaneously considers the spatial and information densities of TFBSs. Prior studies predicting tissue-specific gene expression levels and differentially expressed (DE) TF targets used log likelihood ratios to quantify TFBS strengths and merged adjacent TFBSs into clusters. This thesis presents a machine learning framework that uses the Bray-Curtis function to quantify the similarity between

tissue-wide expression profiles of genes, and IDBC-identified clusters from iPWM-detected TFBSs to predict gene expression profiles and DE direct TF targets. Multiple clusters enable gene expression to be robust against TFBS mutations.

Keywords

Transcription factor binding sites, Shannon information theory, position weight matrices, chromatin immunoprecipitation-sequencing, motif discovery, genetic variants, single nucleotide polymorphisms, mutation analyses, binding site clusters, transcription factor target genes, gene expression, machine learning, Bray-Curtis similarity

Co-Authorship Statement

Chapter 1 – Ruipeng Lu wrote the introduction and created all figures and tables. Peter Rogan provided helpful comments and feedback.

Chapter 2 – Ruipeng Lu wrote the thesis overview. Peter Rogan provided helpful comments and feedback.

Chapter 3 - Peter Rogan conceived of and directed the study. Eliseos Mucaki compared the resultant iPWMs with previous published motifs, and Ruipeng Lu performed other research and analyses. Ruipeng Lu, Eliseos Mucaki and Peter Rogan wrote the manuscript.

Chapters 4 and 5 – Peter Rogan designed, coordinated, and supervised the study, which was motivated by discussions with Joan Knoll regarding prioritization of variants of uncertain significance. Eliseos Mucaki performed probe design and synthesis. Eliseos Mucaki and Natasha Caminsky performed sample preparation and sequencing. Eliseos Mucaki wrote software and performed bioinformatic analysis. Eliseos Mucaki, Natasha Caminsky, and Amy Perri conducted variant analysis and prioritization. Ruipeng Lu generated the TF iPWMs and Eliseos Mucaki generated the RNA-binding protein, splicing regulatory factor, and splicing iPWMs. Ami Perri confirmed prioritized variants by Sanger sequencing. In Chapter 3, Matthew Halvorsen and Alain Laederach conducted the SHAPE analysis. Eliseos Mucaki, Natasha Caminsky, Ami Perri, Joan Knoll, and Peter Rogan wrote the manuscript, which has been approved by all authors.

Chapter 6 - Peter Rogan defined the objectives and directed the study. Ruipeng Lu and Peter Rogan devised the general machine learning framework. Ruipeng Lu implemented this framework and collected the results. Both Ruipeng Lu and Peter Rogan interpreted the results and wrote the manuscript.

Chapter 7 – Ruipeng Lu wrote the discussion and created all tables and figures with helpful comments and feedback from Peter Rogan.

Acknowledgments

First, I would like to thank my supervisor, Dr. Peter Rogan. Four years ago, he gave me this precious opportunity to pursue a PhD degree, so that I was able to fulfill my dream to continue conducting academic research and making contributions to science. During the past four years, he has been wholeheartedly guiding, mentoring, supporting and encouraging me through my research projects. I learned a lot from his diligent research schedule, meticulously scholarly attitude and creatively divergent thinking. Without Dr. Rogan, I would not have anything I have now here. I would also like to thank Dr. Joan Knoll, for her continuous care and encouragement about my research progress.

I would also like to thank Dr. Robert Mercer, Dr. Daniel Lizotte, Dr. Ilka Heinemann, and Dr. Wyeth Wasserman for examining my thesis. I would also like to thank Dr. Robert Mercer for examining my PhD Research Topics Survey/Proposal two years ago and the knowledge I learned from his Computational Linguistics course.

I would also like to thank all of the past and present Rogan/Knoll Lab members. John taught me some basic bioinformatic knowledge after I joined the lab, and I have been happily collaborating with him and Ben on research projects. Li, Ben, Stephanie and Wahab selflessly helped me with my personal issues and research, and I also happily collaborated with Natasha and Amy during the past four years.

Finally, I would like to thank my parents.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1

1 Background

In this chapter, we will introduce preliminary knowledge necessary for understanding our studies described in this thesis, and review previous relevant studies in the literature.

1.1 Transcription Factors

Transcription factors (TFs) are a class of proteins that interact with regulatory elements in genes to facilitate or repress transcription (1). There are two types of TFs, DNA-binding ones and non-DNA-binding ones (2). DNA-binding TFs recognize specific sequence motifs and physically contact these binding sites. Non-DNA-binding TFs form complexes with those sequence-specific TFs as interacting cofactors and are indirectly recruited to regulatory sequences (Figure 1.1). Approximately 2,000-3,000 sequence-specific DNA-binding TFs are estimated to be encoded in the human genome (3).

Sequence-specific TFs can be further divided into two types, depending on whether the binding sites recognized are contiguous or bipartite (4) (Figure 1.1). A contiguous (or single-block) binding site, within which no gap (or spacer) is present, is bound by an individual TF protein (e.g. CAGCTG bound by the TF AP4). A bipartite binding site consists of a left half site, a right half site and a variable-length gap between the two half sites (e.g. TGANTCA/TGANNTCA bound by the TF AP1, where N stands for any base). It is bound by either a homodimer formed by two identical TF subunits, or a heterodimer formed by two different subunits. Among all possible gap lengths, the one which the largest number of binding sites have is referred to as the dominant length.

In general, the binding site motifs of a TF within the entire genome highly resemble each other. However, significant variability in the bases appearing at most positions is present as well. For example, all four sequences (CAGCTG, CACCTG, AACCTG, ATGCTG) are true binding sequences of the TF AP4. TFs exhibit different levels of affinities in their physical association with these different sequences. For example, among these four binding sequences, AP4 has the highest affinity to CAGCTG and the lowest affinity to

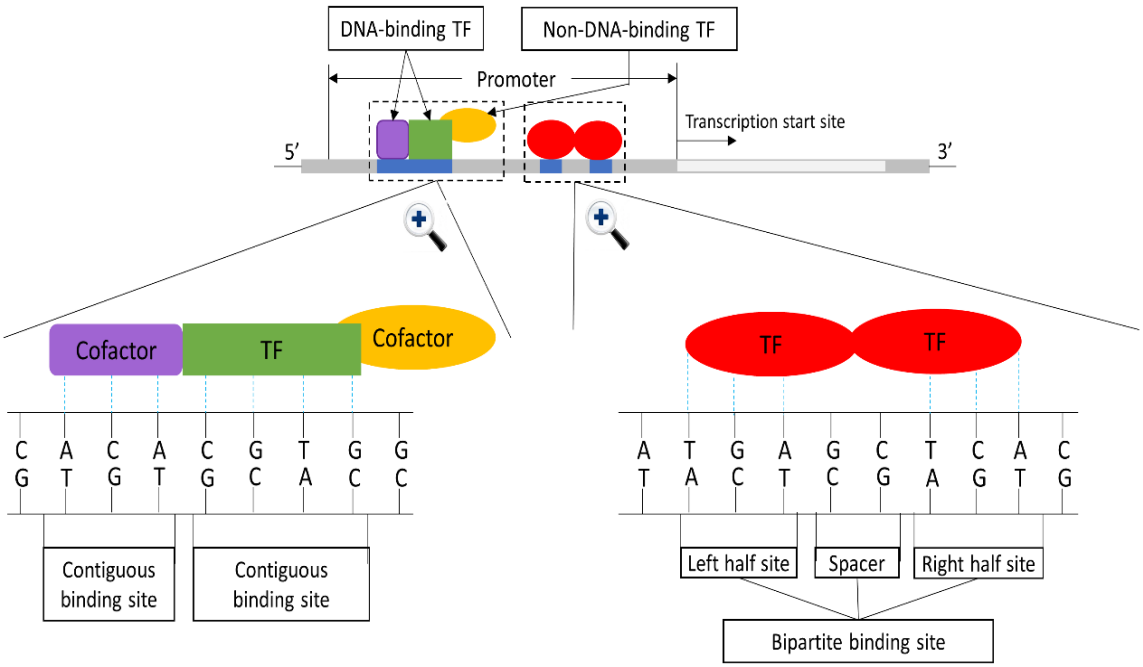

Figure 1.1: Transcription factors. DNA-binding TFs physically associate with binding sites with specific sequence motifs, whereas non-DNA-binding TFs can only be indirectly recruited by forming complexes with sequence-specific TFs. A contiguous TF binding site (TFBS) is recognized by an individual protein, whereas in a bipartite binding site the left and right half sites are respectively recognized by the two subunits of a TF dimer. There also exists coordinate cobinding between sequence-specific TFs. The association of TFs with their binding sites in the promoter effectively regulates the gene transcription rate and expression level.

AACCTG; thus we say CAGCTG are a stronger binding site than AACCTG for AP4. The strongest binding sequence of a TF is referred to as the consensus sequence (e.g. CAGCTG is the consensus sequence of AP4).

Apart from the interplay between the two types of TFs, interactions between sequence-specific TFs also abound across the whole genome (2, 5), which results in the close proximity of their binding sites (Figure 1.1). For instance, NF-Y extensively coassociates with FOS over all chromatin states, and CTCF extensively colocalizes with cohesins consisting of SMC1/SMC3 heterodimers and two non-SMC subunits RAD21 and SCC3 (6, 7).

Functionally, the three-dimensional structure of a TF protein can be divided into multiple domains, including the DNA-binding domain (DBD) and the trans-activating domain (TAD) (8). The DBD is responsible for recognizing and physically associating with specific sequence motifs. The TAD is responsible for forming complexes with interacting cofactors.

TFs can be grouped into major families, based on their structural similarity and behavioral cooperation. For example, the three TFs (FOXA1, FOXA2, FOXA3) belong to the FOXA family, since they share the same DBD named the Forkhead box within their three-dimensional protein structure. And the JUN, FOS and ATF subfamilies belong to the AP1 family, since their members dimerize with each other (e.g. FOS-JUNB and JUND-ATF1) to recognize the bipartite AP1 binding sites with the consensus sequence TGANTCA. Additionally, many TFs are mainly expressed and thus play a role in specific tissues. For example, the FOXA family are predominantly expressed in liver.

Some TFs only act as activators (e.g. SP1 and GATA1) or repressors (e.g. ETV6 and PRDM1); that is, activating TFs can only facilitate the transcription of target genes, and repressing TFs can only impede it. However, there are also a number of TFs that can exert bidirectional effects, partly depending on the interacting cofactors. For example, YY1 induces expression of the NDUSF8 gene encoding the mitochondrial complex I by forming a complex with SP1 (9), whereas YY1 acetylation mediated by the P300 cofactor leads to its repressor activity (10).

1.1.1 Determinants of Transcription Factor Binding

Apart from the core binding sequence physically contacting the TF protein, a number of other determinants can also affect the DNA-binding ability and affinity of TFs, including the sequences flanking the core binding site, local DNA structure and modifications, histone modification patterns surrounding the binding site, chemical modifications within the TF protein itself, interacting cofactors and ligand signals, and the spacer sequences within bipartite binding sites (Table 1.1).

Table 1.1: Determinants affecting TF binding besides the core binding sequence

Determinant	Evidence	Reference
Flanking sequences	The base pairs flanking an EGR1 binding site in the LHB promoter modulates the affinity and structure of the protein-DNA complex.	(11)
	The nucleotides flanking GR binding sites change the 3D structure of the binding site, the DNA-binding domain of GR and the quaternary structure of the dimeric complex.	(12)
	Zinc finger TFs of C2H2 type prefer GC-rich over the AT-rich flanking sequences.	(13)
DNA shape	The A-DNA structure is only present in DNA bound to TFs and avoided in DNA bound to nucleosomes, whereas the BII-DNA structure periodically occurs every 10.3 dinucleotide steps in DNA bound to nucleosomes.	(14)
	Both global DNA shape (e.g. an overall bend) and local shape (e.g. a kinked base pair or a narrow minor groove) determine the TF binding specificity.	(15)
	The inherent deformability of the TATA sequence assists in TBP to distinguish bound from unbound sites.	(16)
DNA modifications	Promoter methylation levels of the MGMT gene in the glioma cell line correlates with chromatin accessibility and SP1-DNA interaction levels.	(17)
Histone modifications	HM patterns surrounding binding sites differ considerably from those surrounding non-sites in a TF family-specific manner.	(18)

TF protein modifications	Phorbol ester-induced dephosphorylation of JUN strongly increases its DNA-binding potential.	(19)
Cofactors	NR3C1 (GR) activated by glucocorticoid complex with NFKB, AP1, T-bet and GATA3 to inhibit their DNA-binding ability.	(20–23)
Ligand signals	Association with glucocorticoid hormone activates NR3C1, enabling it to directly bind the response element or complex with other TFs.	(24)
Spacer sequences (for dimer TFs)	A single pyrimidine nucleotide at the third position of the spacer in the TR/RXR bipartite site enhanced its binding and transactivation.	(25)

The 3-5 nucleotide (nt) sequences flanking the core binding site can profoundly affect TF binding by altering the DNA shape, TF DBDs, the quaternary structure of dimeric TFs, and possibly TF search dynamics (34, 48, 49). The effects of the flanking sequences can even be asymmetric. For example, a binding site of EGR1 harboring three zinc fingers (ZF1, ZF2, ZF3) is present in the LHB gene promoter (11). Base substitutions in the sequence flanking ZF3 resulted in a more significant change in the binding site strength (11). In addition, the similarity of the flanking sequence composition positively correlates with the core binding site strength. For instance, the C2H2-type zinc finger DBD recognize GC-rich sequence motifs, consistent with the finding that the TFs with this DBD prefer GC-rich over the AT-rich flanking sequences (13).

Individual TF proteins combine two readout mechanisms, recognition of a unique DNA base sequence (base readout) and of a sequence-dependent DNA shape (shape readout) to achieve DNA-binding specificity (15). The specificities of different TF families differ in the base readout in the major groove, whereas shape readout distinguishes between members within a family (15). Promoter methylation levels of the MGMT gene was also found to be related to TF binding levels by altering chromatin accessibility (17). Within open chromatin, different TF families also exhibited different preferences for HM

patterns surrounding binding sites, with high consistency across cell lines (18). Phorbol esters, a tumor-promoting agent, induced the dephosphorylation of JUN proteins which dimerize with other members of the AP1 TF family. This subsequently increased the DNA-binding activity of the TF to AP1 bipartite sites (19). Association with other cofactors can also alter TF binding, e.g. NFKB, AP1, T-bet and GATA3 are inhibited by complexing with activated NR3C1 (20–23). Hormone receptors (e.g. NR3C1) can only be activated by binding to the corresponding hormone ligands (24).

As described above, for dimeric TFs, the spacer length between the two half sites affects the TF affinity to the bipartite binding site, with the dominant spacer length being the most favored (4). In addition, the spacer sequence also modulates the binding site strength and further the activation potency that the TF exerts on gene transcription. For example, between the directly repeated half sites with a 4-nucleotide (4nt) spacer recognized by the heterodimer of TR (thyroid hormone receptor) and RXR (retinoid-X receptor), some spacer sequences allowed little or no transactivation, whereas other sequences supported strong transactivation (25). Specifically, a single pyrimidine nucleotide at the third position of the spacer enhanced TR/RXR binding and transactivation (25). Heterodimers between RXR and other receptors exhibited a similar but distinct specificity for the spacer sequence (25).

Most TFs are unable to bind DNA within closed chromatin; however, there is one category, pioneer TFs (e.g. FOXA1, POU2F2, PAX7 and SPI1), that can be the first to target closed chromatin and engage binding sites (26). Such initial binding enhances transcription by reducing the number of additional factors that are needed to bind the DNA, and actively opening up the local chromatin to enable other factors to bind (26–28). On the other hand, pioneer factor binding can also lead to repressed chromatin, through binding adjacent to repressors or corepressors and reduced local nuclease sensitivity (29–31).

1.1.2 Impacts of Transcription Factor Binding

As described above, TF binding to target sites can eventually result in the effective regulation of the gene transcription rate and expression level; this is achieved first via

direct physical impacts that TF binding causes on DNA conformation and chromatin structure, including DNA bending, over-twisting and/or untwisting, and nucleosome displacement (Table 1.2).

Table 1.2: Physical impacts of TF binding on DNA and chromatin

Impact	Evidence	Reference
DNA bending	Within MADS-box proteins, SRF induces considerable DNA bending into its binding sites, whereas MEF2A induces minimal DNA bending.	(32)
	TBP induces linear, stepwise DNA bending with an intermediate state distinguished by a distinct bending angle.	(33, 34)
	U-turn: The human TRAM TF forces mitochondrial promoter DNA to undergo a U-turn, reversing the direction of the DNA helix both *in vivo* and in solution.	(35, 36)
	Looping: Ribosomal promoter DNA is looped into a single 180-base pair turn around the xenopus UBF dimer, probably by in-phase bending.	(37)
	Supercoiling: Topological stresses caused by TF-induced pronounced bending on circular DNA are compensated by DNA supercoiling.	(38)
DNA over-twisting and untwisting	The 434 repressor overtwists its binding site upon DNA binding.	(39)
	The zinc finger DBDs of SP1, GLI and ZIF268 unwind DNA upon binding.	(40–42)
Nucleosome displacement	TFs compete with nucleosomes for binding DNA to produce nucleosome free regions in promoters.	(43)

DNA bending induced by TFs is thought to be an important facet of their function, and plays a role in the DNA recognition process and determining the correct architecture of nucleoprotein complexes at promoters and enhancers (32). For example, within the TF family sharing the MADS-box DBD, SRF induces considerable DNA bending into its binding sites (44); by contrast, MEF2A induces minimal DNA bending (45). The local DNA architecture surrounding the promoter-bound SRF and MEF2A will therefore differ significantly, and may contribute to their different biological functions (32). The TFIID complex that includes the DNA-binding TBP as the core subunit induces a linear, stepwise bending process with an intermediate state hallmarked by a distinct bending angle around the TATA motif (33, 34).

An extreme case of DNA bending is a U-turn shape where the direction of the DNA helix is completely reversed, such as the one induced by the TF TRAM on the mitochondrial promoters (35, 36). If the two ends of a U-turn bending intersect, then a loop, maybe around TF proteins themselves (37), will form. On a circular DNA molecule, topological strain caused by sufficiently strong bending will be balanced by supercoiling, which can be seen as another higher-level double helix besides the inherent double helix formed by the two DNA strands (38).

DNA untwisting resulting from binding of TFs with DBDs of zinc finger type (e.g. SP1, ZIF268, GLI) may affect binding site affinities and TF-cofactor interactions (40–42); by contrast, TF binding can also overtwist binding sites, such as the bacteriophage 434 repressor (39). In addition, TFs also compete with nucleosomes for binding DNA (43). Larger nucleosome free regions in promoters, which likely are open chromatin and have a much more significant impact on gene expression, are determined mainly by TF binding (43).

1.2 Information Theory-based Position Weight Matrices

Since the bases appearing at most positions of the binding site of a TF are highly variable, the single consensus sequence is not able to accurately represent the binding specificity of the TF by only indicating the most frequent base at each position. In contrast, a position weight matrix (PWM) can more accurately describe the base

preference of the TF by accounting for the conservation level of each base at each position. It is a commonly used representation of motifs in biological sequences (46). It has one column for each position in the motif, and one row for each symbol of the alphabet: 4 rows for nucleotides in DNA sequences or 20 rows for amino acids in protein sequences.

1.2.1 Derivation of Information Theory-based Position Weight Matrices

Contiguous and bipartite information theory-based position weight matrices (iPWMs) respectively can quantitatively describe the base preferences of TFs recognizing contiguous and bipartite binding sites.

A contiguous iPWM is derived from a set of aligned binding sites based on Shannon information theory (4, 47) (Figure 1.2). From a multiple alignment of n binding sites in the reference genome (48), the iPWM is computed from

$$R_{iw}(b,l) = 2 - \left(-\log_2 f(b,l) + e\big(n(l)\big)\right)(bits\ per\ base) \quad [1.1]$$

where $f(b,l)$ is the frequency of base b at position l in the alignment (i.e. the count of b at l divided by n), and $e\big(n(l)\big)$ is a sampling error correction factor (49) at position l for the n sequences used to create $f(b,l)$ (50, 51). $e\big(n(l)\big)$ exists, since this set of binding sites does not include all possible binding sites in the genome and using sampling frequencies in place of population probabilities leads to a bias in the uncertainty measurement (49, 52). It is approximately computed from

$$e\big(n(l)\big) = \frac{s-1}{2\ln(2)n} \quad [1.2]$$

where s, the number of symbols, is 4 for nucleotides (49, 52). It is inaccurate for small n values but accurate for large n values (49). The exact method is given in Appendix A.2.

In the two-dimensional matrix $R_{iw}(b,l)$, row b corresponds to one of the four nucleotides in DNA and column l is the position along the aligned binding sites (50). Following

Shannon's convention, R_{iw} stands for "Rate of information transmission, Individual Weight" (51). Bits per base is a rate like bits per second, especially if we consider the average binding rate in bases per second (49). This individual information matrix represents the sequence conservation of each nucleotide, measured in bits of information (50). A bipartite iPWM consists of two contiguous iPWMs, each of which corresponds to a half site, separated by a range of sequence gaps with penalties (4).

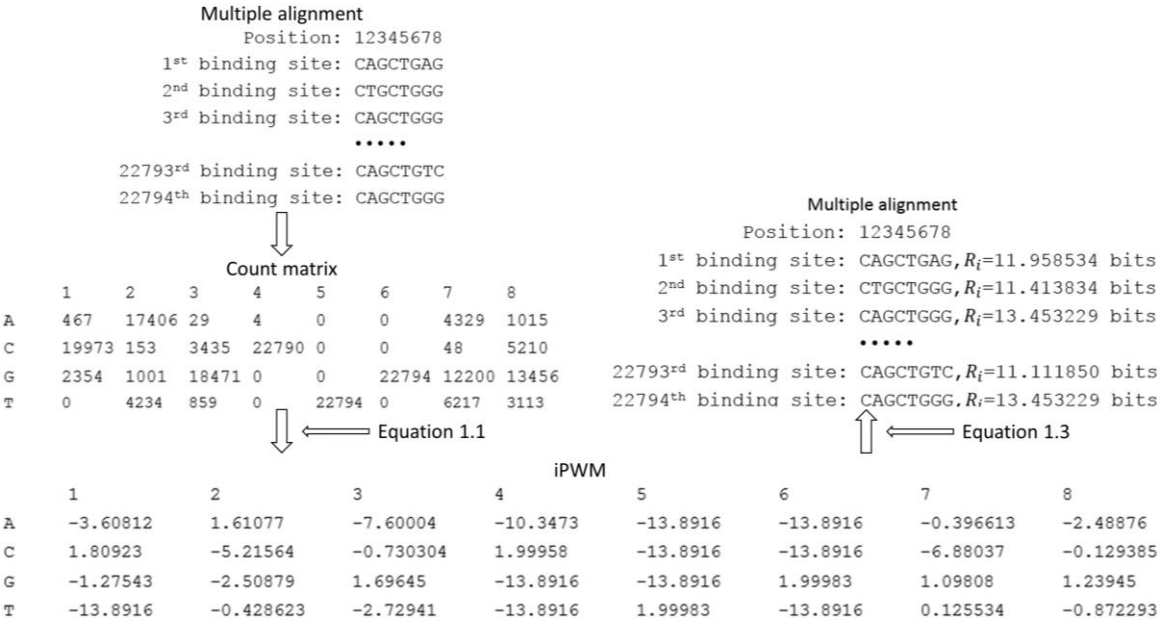

Figure 1.2: Derivation of a contiguous iPWM from a multiple alignment of binding sites. An aligned set of 22,794 binding sites of the TF AP4 is shown. A count matrix is obtained from the alignment by counting the occurrence of each base at each position. The information theory-based Equation 1.1 converts the count matrix to an iPWM using a pseudocount 0.375 for each base (4). Equation 1.3 uses the iPWM to compute the $\boldsymbol{R_i}$ values of all binding sites in the alignment from which the $\boldsymbol{R_{sequence}}$ value is computed.

1.2.2 Computation of the Information Content of an Individual Binding Site

The individual information content of a contiguous binding site sequence j, which represents its strength, can be determined using a contiguous iPWM. It is the dot product between the sequence and the iPWM:

$$R_i(j) = \sum_{l} \sum_{b=A}^{T} s(b,l,j) R_{iw}(b,l) \ \ (bits\ per\ site) \quad [1.3]$$

where $s(b,l,j)$ is a binary matrix for the sequence j, in which cells have a value of 1 for base b at position l and a value of 0 elsewhere (50, 51) (Tables 1.3 and 1.4).

Table 1.3: The binary matrix of the binding site sequence "CATCTGGG" of AP4

Base b	Position l^*							
	C	A	T	C	T	G	G	G
	1	2	3	4	5	6	7	8
A	0	1	0	0	0	0	0	0
C	1	0	0	1	0	0	0	0
G	0	0	0	0	0	1	1	1
T	0	0	1	0	1	0	0	0

*There is only one "1" in each column, indicating the base appearing at that position.

Table 1.4: Calculation of the R_i value of the binding site sequence "CATCTGGG"

Base b	Position l^*							
	C	A	T	C	T	G	G	G
	1	2	3	4	5	6	7	8
A	-3.60812	**1.61077**	-7.60004	-10.3473	-13.8916	-13.8916	-0.396613	-2.48876
C	**1.80923**	-5.21564	-0.730304	**1.99958**	-13.8916	-13.8916	-6.88037	-0.129385
G	-1.27543	-2.50879	1.69645	-13.8916	-13.8916	**1.99983**	**1.09808**	**1.23945**
T	-13.8916	-0.428623	**-2.72941**	-13.8916	**1.99983**	1.99983	0.125534	-0.872293

*In this iPWM, the individual weights selected by the binary matrix of the binding site are bolded. In Equation 1.3 the R_i value is the sum of the selected weights.

The R_i value of a bipartite binding site sequence k is the sum of the R_i values of the two half sites, each of which is computed from Equation 1.3, subtracting the gap penalty (4):

$$R_i(k) = R_i(k_l) + R_i(k_r) - g(d) \quad [1.4]$$

where k_l and k_r are the two half sites of k, d is the gap length between k_l and k_r, and $g(d)$ is the function used to compute gap penalties:

$$g(d) = 1 - \log_2(1 + \cos(2\pi(d-c)/B)) \quad [1.5]$$

where c is the dominant gap length (i.e. the number of binding sites with the gap length c in the alignment from which the bipartite iPWM is computed is the largest), and B is a DNA helical turn (10.4 bases/turn) (4, 53). Equation 1.5 incorporates the geometry of the TF recognition to binding sites, that is, the preference of dimeric TFs for binding across adjacent major grooves in DNA helices (53).

1.2.3 Relationship between R_i Values and Thermodynamics

According to the Second Law of Thermodynamics, the relationship between information (i.e. R_i values) and the heat q:

$$k_B T \ln(2) \leq \frac{-q}{R_i} \quad [1.6]$$

where k_B is Boltzmann's constant (1.38×10^{23} joules/K), T is the absolute temperature in Kelvin.

Equation 1.6 implies that the association process of a TF protein with a binding site ($R_i > 0$) dissipates heat energy ($q < 0$), and the association process of a TF protein with a non-site ($R_i < 0$) absorbs heat energy ($q > 0$). This suggests that TFs have a natural tendency to complex with its binding sites, and the association of TFs with non-sites does not naturally occur. Furthermore, Equation 1.6 implies that if a SNP results in a 1-bit increase in the R_i value of a binding site, the amount of heat energy dissipated by the association of a TF protein with this site will at least increase by $k_B T \ln(2)$. The derivation starting from the Second Law of Thermodynamics to obtain Equation 1.6 is in Appendix A.1.

Thus based on the R_i value, a nucleotide sequence can be predicted to be a binding site or not; if $R_i > 0$ it is a binding site, and if $R_i < 0$ it is a non-site. The $R_{sequence}$ value of an iPWM is the mean of the R_i values of all binding sites used to compute the iPWM, representing the average binding strength of the TF (50, 51). The distribution of R_i values is approximately Gaussian, with the mean being $R_{sequence}$; however, the lower and upper bounds are zero bits and the R_i value of the consensus sequence (50, 51).

The change in the binding site strength caused by a SNP can be quantitatively computed from the change (ΔR_i) in the R_i value of the binding site (50, 51) (Appendix A.3). If the R_i values of the normal and variant binding sites are respectively $R_{i,normal}$ and $R_{i,variant}$, then $\Delta R_i = R_{i,variant} - R_{i,normal}$. Then the minimum fold change between the binding affinity of the two sites is $2^{\Delta R_i}$ (50, 51). R_i and ΔR_i will be used to detect experimentally confirmed binding sites and interpret experimentally measured effects of SNPs in Chapter 3, to detect and prioritize TFBS variants in Chapters 4 and 5, and to perform mutation analyses on promoters of TF target genes in Chapter 6.

1.2.4 Sequence Logos

An iPWM can quantitatively describe the binding specificity of a TF, but it is not very intuitive. Sequence logos provide a means to intuitively visualize iPWMs in Chapter 3 (Figure 1.3) (54). In a sequence logo, the abscissa is the position along the aligned binding sites; the ordinate is the individual weight in the iPWM, so that the height of a base letter indicates the conservation level of the base at the position among binding sites.

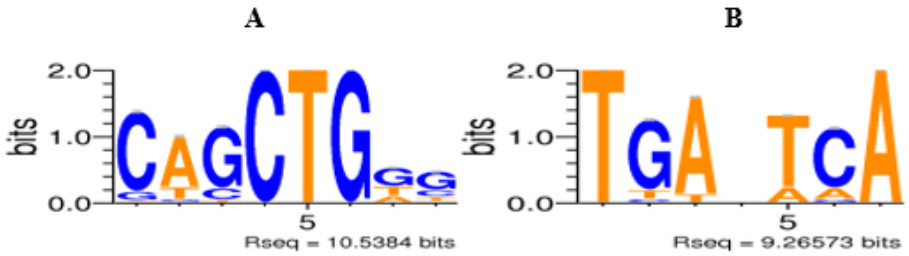

Figure 1.3: Sequence logos. The height of each base letter is its individual weight at the position in the iPWM. The $R_{sequence}$ value of the iPWM is indicated at the bottom of its logo. (A) This logo visualizes the contiguous iPWM of AP4 derived in Figure 1.5. (B) This logo visualizes a bipartite iPWM of the TF BATF. From this logo, we can see that the length of both half sites is 3bps and the dominant gap length is 1bp.

1.3 Derivation of Transcription Factor Binding Site Motifs

As described above, binding site sequences recognized by a TF are highly variable; thus one question is how we can identify all these true binding site sequences. Previous studies, using either *in vitro* experimental techniques or computational approaches based

on *in vivo* generated ChIP-seq datasets, have derived TF binding site sequence motifs and characterized TF binding specificities. Compared to experimental techniques, the computational approaches are less labor intensive and more cost effective.

1.3.1 Using Experimental Techniques

In Weirauch et al. (55), binding site motifs of more than 1,000 TFs belonging to 54 different DBD classes from 131 eukaryotes were derived using PBMs (protein binding microarrays) to determine their sequence preferences. In each PBM, each possible 8-mer is present 32 times, allowing for a robust and unbiased assessment of TF binding affinity. Since closely related DBDs have similar sequence preferences, they were further able to infer motifs for 34% of the approximately 170,000 known or predicted eukaryotic TFs (55). These binding site motifs were validated by the fact that they are enriched in ChIP-seq peaks and upstream of TSSs (55). Their results, in the form of frequency matrices, were stored in the CIS-BP database (55).

In Jolma et al. (56), HT-SELEX (high-throughput systematic evolution of ligands by exponential enrichment) experiments were performed to identify binding site motifs for 151 human full-length TFs and 303 human DBDs, and PWMs were further generated using a multinomial model. Pairwise comparison between the results of full-length TFs and their DBDs revealed that the sequence preference of a TF is primarily determined by its DBD (56). The vast majority of physical interactions between a TF and individual DNA bases were found to be independent of each other (56). Bipartite binding site motifs of TF homodimers with strong orientation and spacing preferences were also observed based on the presence of two similar sites in a single oligonucleotide (56).

However, these oligonucleotide-based approaches have a number of limitations:

1) Weirauch et al. did not derive bipartite binding sites of dimeric TFs from their PBM experiments; Jolma et al. only generated contiguous PWMs with dominant gap lengths to describe bipartite motifs of homodimers, which are unable to account for the variability of gap lengths.

2) Since these techniques use oligonucleotides whose lengths are fixed and limited, they are unable to completely derive motifs of TFs whose binding site lengths exceed the oligonucleotide lengths.

3) There is no way to discover potential binding sites of interacting cofactors, since only the primary TF proteins of interest are added to each experiment.

1.3.2 Using Computational Approaches

1.3.2.1 Chromatin Immunoprecipitation-Sequencing Assays

ChIP-seq, which combines chromatin immunoprecipitation with massively parallel DNA sequencing, is an *in vivo* experimental method to identify the genome-wide repertoire of binding sites of chromatin-associated proteins (e.g. TFs) and the distribution of histone modifications (57) (Figure 1.4).

After cross-linking proteins to DNA in living cells, DNA molecules are sheared into short fragments by sonicating. Using an antibody against the protein of interest, chromatin immunoprecipitation (ChIP) isolates the DNA-protein complexes, resulting in a library of DNA fragments directly bound to the protein (Figure 1.4). Then these fragments are sequenced and mapped to the reference genome, ultimately producing a dataset file containing genome coordinate intervals (Figure 1.4). In this dataset, each interval, typically several hundred nucleotides long, has a signal value that indicates the enrichment level of mapped DNA fragments within this interval; thus in general, the signal value is positively correlated with the strength of the binding site contained in the interval. Each interval is referred to as a peak. The process of mapping DNA fragments to the genome, merging DNA fragments into peaks based on their local enrichment levels, and determining the signal values of these peaks is referred to as peak calling. In the case of TFs, the DNA fragments in the library also contain binding sites of cofactors, due to either the proximity between their binding sites resulting from the TF-cofactor coordination or the recruitment of non-DNA-binding TFs by sequence-specific cofactors.

The ENCODE (Encyclopedia of DNA Elements) Consortium conducted ChIP-seq assays for human TFs and histone modifications, and generated an initial peak dataset for each

replicate of each assay using a uniform peak calling pipeline (57, 58). To improve the consistency between the initial peaks from multiple replicates of the same assay, for some assays it also produced optimal and conservative IDR (irreproducible discovery rate)-thresholded peaks after applying the IDR framework to the initial datasets (59). In addition, members of the ENCODE Consortium also individually generated refined datasets using the SPP peak calling software (60).

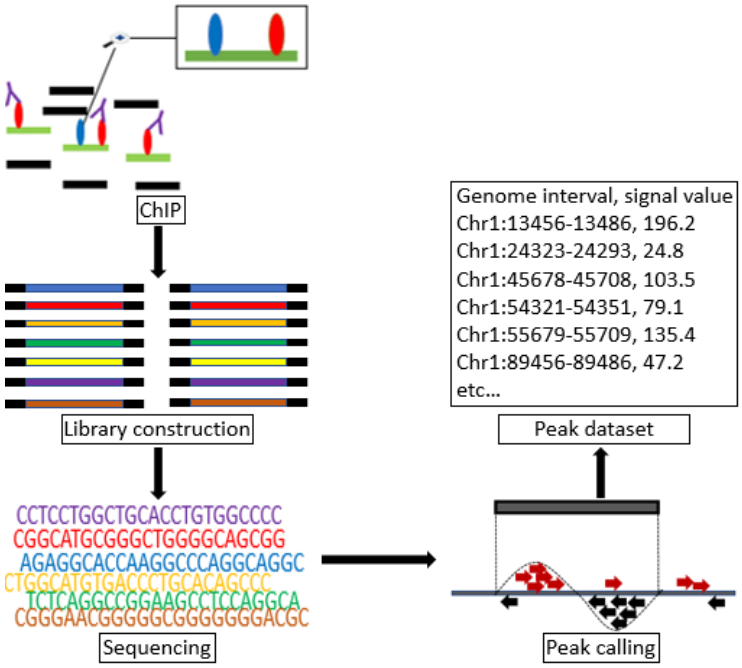

Figure 1.4: ChIP-seq assays. The protein binding to DNA fragments *in vivo* is immunoprecipitated by an antibody. These fragments that frequently contain binding sites of cofactors and form a library containing a genome-wide set of binding sites of the protein, are sequenced. The peak calling algorithm identifies all intervals (or peaks) with high enrichment of the DNA fragments and computes associated signal values. The genome intervals and signal values are stored in a peak dataset file.

1.3.2.2 Motif Discovery Algorithms

A peak in a ChIP-seq dataset contains a binding site sequence of the TF; however, it also contains long unnecessary sequences flanking the binding site at the same time. Previous studies have developed multiple algorithms and corresponding software programs to identify the accurate binding site motifs from ChIP-seq datasets (Table 1.5). These motif discovery algorithms fall into six categories, depending on the mathematical principle

Table 1.5: Motif discovery algorithms

Algorithm†	Mathematical principle	Ability to derive bipartite motifs	Reference
Bipad	Entropy minimization	√	(4)
MEME /MEME-ChIP	Expectation maximization	×	(61–63)
SeqGL	Group lasso regularization	×	(64)
MotifCut	Maximum density subgraph	×	(65)
AlignACE	Gibbs sampling	×	(66, 67)
ANN-Spec		×	(68)
GLAM		×	(69)
GLAM2		√	(70)
MotifSampler		×	(71)
SeSiMCMC		√	(72)
MDscan	String matching	×	(73)
Trawler		×	(74)
Weeder		×	(75)
MITRA		×	(76)
DREME		×	(77)
HOMER		×	(78)
YMF		×	(79)

† The rows of the motif discovery algorithms using the same mathematical principle have the same background shade.

used, including entropy minimization (used by Bipad), expectation maximization (used by MEME), group lasso regularization (used by SeqGL), maximum density subgraphs (used by MotifCut), Gibbs sampling (used by AlignACE, ANN-Spec, GLAM/GLAM2,

MotifSampler, SeSiMCMC), string matching-based enumeration (used by MDscan, Trawler, Weeder, MITRA, DREME, HOMER, YMF).

These algorithms were benchmarked on 52 datasets, finding that Weeder outperformed others on most datasets (80); however, Bipad was excluded from the comparison. Among these algorithms, only three (i.e. Bipad, GLAM2 and SeSiMCMC) are capable of deriving bipartite motifs. Both MEME and SeqGL are discriminative methods that distinguish motif from background DNA in a mathematically optimal way using background nucleotide frequencies computed from all input sequences (60, 61, 63).

1.3.2.2.1 The Bipartite Pattern Discovery Algorithm

The Bipartite Pattern Discovery (Bipad) algorithm, that uses an entropy minimization-based Monte Carlo framework, can derive both contiguous and bipartite iPWMs from the multiple alignment of binding sites identified from a ChIP-seq dataset (4). Chapter 3 will improve its ability to discover known TFBS motifs and use it to analyze ENCODE ChIP-seq datasets.

The Bipad algorithm assumes that the sequence associated with each peak in a ChIP-seq dataset contains one binding site. Given the length J of the contiguous binding sites of the TF and the peak count n in the dataset, all peaks form a multiple alignment search space Θ. In Θ each multiple alignment is derived by extracting a sequence fragment of length J from each peak and aligning them; thus it is of width J and size n. Given a multiple alignment MA in Θ, the entropy of the position l is computed from:

$$H_l = \sum_{b \in B} f(b,l) \log_2 \frac{1}{f(b,l)}, B = \{A, C, G, T\} \quad [1.7]$$

where $f(b,l)$ is the frequency of base b at position l. The entropy of the alignment MA is computed from:

$$H_{MA} = \sum_{l=1}^{J} H_l \quad [1.8]$$

Similarly, in the instance of bipartite binding sites of a dimeric TF, the entropy of the

bipartite alignment MA is computed from:

$$H_{MA} = \sum_{s\in\{L,R\}} \left(\sum_{l=1}^{J_s} H_l \right) \quad [1.9]$$

where J_L and J_R are respectively the lengths of the left and right half sites.

The Bipad algorithm uses multiple Monte Carlo cycles to search Θ for the optimal alignment with the minimum entropy (Figures 1.5 and Appendix A.4). Its objective function is:

$$oMA = arg \min_{MA\in\theta} (H_{MA}) \quad [1.10]$$

where oMA is the optimal alignment with the minimum entropy, and MA is a contiguous or bipartite alignment in Θ. Its time complexity and space complexity are respectively $O(JLnc)$ and $O(Ln)$, where L is the length of a ChIP-seq peak and c is the Monte Carlo cycle count.

At the beginning of each cycle, the initial position of the binding site in each peak, and thus the initial alignment, are randomly generated. Each cycle performs multiple iterations; in each iteration, the binding site in each peak moves to every possible position (including every possible gap length at every position in the instance of bipartite binding sites) to generate a new alignment. The end of each cycle converges to a locally optimal alignment in a subspace of θ.

1.3.2.3 Application of the Motif Discovery Algorithms to ChIP-seq Datasets

These motif discovery algorithms have been applied by prior studies to ChIP-seq datasets. For example, Wang et al. performed *de novo* motif discovery from top 500 peaks with the highest signal values of 457 ChIP-seq datasets of 119 human TFs using the MEME-ChIP software (2, 63). Apart from known and new motifs of primary TFs, they also derived cofactor motifs adjacent to them, indicating tethered binding and cobinding between multiple TFs (2). Specifically, they observed cell line-specific

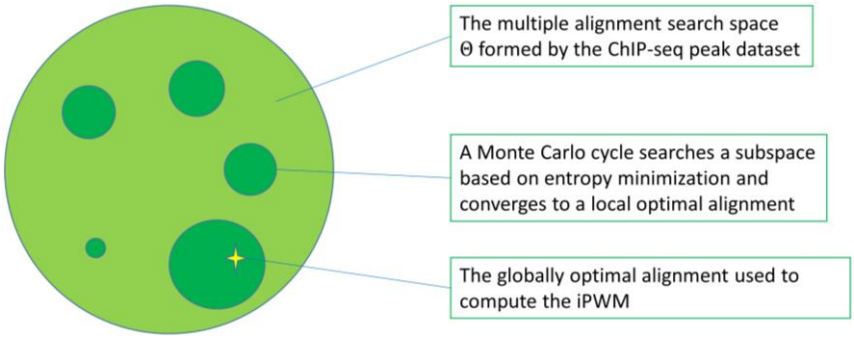

Figure 1.5: An execution of the Bipad algorithm on a ChIP-seq dataset. Using multiple entropy minimization-based Monte Carlo cycles, the Bipad algorithm searches the multiple alignment search space formed by all peaks, in an attempt to find the optimal alignment with the minimum entropy.

cofactor motifs that mediate the binding of the histone deacetylase HDAC2 and the enhancer-binding protein EP300 (2). They created the Factorbook database to make the derived binding site motifs publicly available (2). Kheradpour et al. also confirmed known binding motifs and revealed potential cofactors from top 250 peaks of 427 ChIP-seq datasets of 123 TFs using five motif discovery tools (AlignACE, MDscan, MEME, Trawler, Weeder) (81).

Setty et al. applied SeqGL to top 2,000 peaks of 105 ChIP-seq datasets from the GM12878 and H1-hESC cell lines, and found that it outperformed three other widely used motif discovery algorithms (HOMER, DREME and MEME-ChIP) in terms of both locating motifs of primary TFs and revealing new cell type-specific cofactors (64). In addition, SeqGL also successfully detected binding site motifs from DNase-seq and ATAC-seq datasets (64).

These previous studies mined a wealth of valuable information in terms of TF binding specificities and TF-cofactor interactions. However, they also have a number of limitations:

1) They only derived contiguous binding site motifs, and did not generate bipartite motifs with variable-length gaps for dimeric TFs. Thus the contiguous PWMs they provided are unable to accurately reflect the binding behavior of dimeric TFs or further detect their binding sites.

2) They only respectively analyzed top 500, 250 or 2,000 peaks. Because the signal value of a peak is generally positively correlated with the strength of the binding site contained in the peak, this implies that they only obtained strong binding sites, resulting in the fact that the PWMs only represent strong sites and cannot accurately detect weak sites.

3) The PWMs that they generated to describe the binding site motifs were not Shannon information theory-based. The log likelihood ratios computed from these PWMs to detect TFBSs indicate the probabilities that DNA sequences are binding sites, which are not quantitatively related to the amount of dissipated binding energy by Equation 1.6. Thus these PWMs are unable to quantify binding site strengths as accurately as iPWMs by computing R_i values.

4) Both the MEME and SeqGL algorithms compute background nucleotide frequencies from all input sequences, then use them to discriminate binding site motifs from background sequences. Thus they may fail to discover motifs with compositions similar to the background.

1.3.3 Transcription Factor Binding Site Motif Databases

Multiple databases have been created to store PWMs and sequence logos describing TFBS motifs experimentally or computationally derived by these prior studies (Table 1.6). The PWMs in these databases have been widely used to detect binding sites in previous studies.

The JASPAR database initially only contained 111 pan-species count matrices derived from a limited number of experimentally validated binding sites (82), but the core collection of its latest 2018 version has been significantly expanded to 1404 pan-species non-redundant position frequency matrices (PFMs) by incorporating TFBS motifs computationally derived from ChIP-seq datasets (83). The Factorbook database (84), containing binding site motifs derived from top 500 peaks of ChIP-seq datasets using the MEME-ChIP software (2), was also increased from the initial 119 TFs to 167 TFs. The CIS-BP database contains the frequency matrices derived from the octanucleotide-based PBMs by Weirauch et al. (55). The TRANSFAC database (85) currently contains 7,371

Table 1.6: TFBS motif databases

Database	Human TF/PWM count†	PWM source	Reference
JASPAR (Core collection)	537 PWMs	Experimental (PBM, HT-SELEX, etc) and computational (ChIP-seq)	(83)
Factorbook	167 TFs	Computational (ChIP-seq)	(84)
CIS-BP	972 PWMs, 1734 TFs	PBM experiments	(55)
TRANSFAC	Unavailable	Computational (phylogenetic analysis) and possibly experimental	(85)

† We were not able to obtain accurate both human TF and PWM counts for JASPAR and Factorbook. Neither human TF nor PWM counts could be obtained for TRANSFAC.

pan-species PWMs, a small fraction of which are freely available due to its commercial orientation.

1.4 Transcription Factor Binding Site Variants

One source that can result in misregulation of gene expression is variation. A genetic variation in the human genome is defined as a difference in the DNA sequence between two individuals or paired chromosomes in an individual (86). Multiple types of variants (e.g. nucleotide substitutions, insertions and deletions) can occur within both the non-coding and coding regions of genes, affecting the binding sites of TFs, splicing regulatory proteins and RNA-binding proteins, and protein structure (50, 87, 88). Specifically, genetic variants within exons of the coding region of a gene can alter individual residues in the amino acid chain of the protein or result in a prematurely truncated chain, and variants within splicing sites can also change the polypeptide by altering the strengths of splice sites and potentially the boundaries between introns and exons (50, 87). Thus variants are often associated with the increased risks of a variety of diseases (89, 90). For example, nine splice site variants in the promoter of the ABCR gene are found to be

associated with the onset of the STGD (autosomal recessive Stargardt disease) and AMD (cone-rod dystrophy) (91).

Among variants within TFBSs, a single nucleotide polymorphism (SNP) is the simplest case, which is a naturally occurring substitution of one single nucleotide with a prevalence rate exceeding 1% in a population (92). It can significantly alter the strength of the binding site, and further the gene expression level (Figure 1.6). The expression level of the gene will increase after a SNP strengthens a binding site of an activating TF or weakens a site of a repressing TF; by contrast, it will decrease after a binding site of an activating TF is weakened or a site of a repressing TF is strengthened (Figure 1.6).

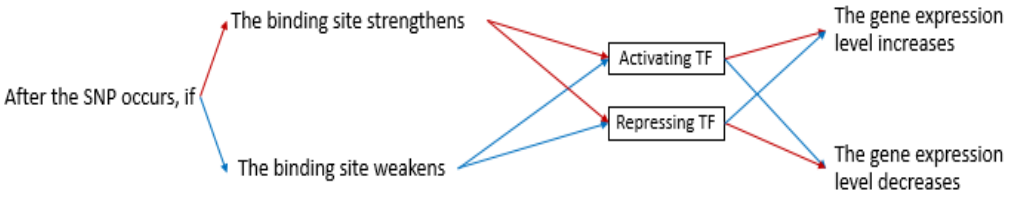

Figure 1.6: The impacts of a SNP on the TFBS and gene expression level. The strengthening or weakening of the binding site of an activating TF caused by a SNP respectively leads to an increase or decrease in the gene expression level. The strengthening or weakening of the binding site of a repressing TF caused by a SNP respectively leads to a decrease or increase in the gene expression level.

PWMs have been used to detect variants within TFBSs and splice sites and predict their effects, either as online web-based services or stand-alone software programs (Table 1.7). The majority of these platforms use the PFMs from the JASPAR database to compute log likelihood ratio scores, except for Shannon pipeline using iPWMs. These scores are not Shannon information theory-based, so that they are unable to quantify binding site strengths as accurately as R_i values.

1.5 Clustering of Transcription Factor Binding Sites

The distinctive organization and combination of transcription factor binding sites (TFBSs) and regulatory modules in the promoters of human genes substantially dictate specific expression patterns within a set of genes (93). Clustering of multiple adjacent

Table 1.7: Platforms applying PWMs to detect TFBS and splice site variants

Platform	PWM Type	PWM source	Service Type	Reference
Shannon pipeline	iPWM	Manually curated splice sites	Web-based	(94)
rSNP-MAPPER	PFM	JASPAR, TRANSFAC		(95)
SNP2TFBS		JASPAR		(96)
OncoCis				(97)
RAVEN				(98)
atSNP		JASPAR, Kheradpour et al.		(99)
TRAP		JASPAR, TRANSFAC	Stand-alone	(100)
motifbreakR		JASPAR, Factorbook		(101)

binding sites for the same TF (homotypic clusters) and for different TFs (heterotypic clusters) defines regulatory modules, and are prevalent in both promoters and more distal enhancers (102). Evolutionarily conserved homotypic clusters occupy nearly 2% of the human genome. The promoters of more than half of human genes contain homotypic clusters, with a concentrated distribution around the TSS (103). For example, highly degenerate binding sites of NRSF, MYC, p53, HNF1 and CREB were found to have a tendency of non-randomly clustering around the cognate binding sites (104).

1.5.1.1 Impacts of Transcription Factor Binding Sites Clusters

There are two scenarios under which a homotypic TFBS cluster can influence gene expression, depending on whether individual binding sites within the cluster interact with one another (102).

1.5.1.1.1 Absence of Interaction Between Individual Sites

If there is no interaction between individual sites at all, all binding sites in a cluster are equally likely to be bound, and the likelihood may be associated with an external variable

to the system, such as TF concentration (102). This scenario can be further divided into the following four different cases (here we assume an activating TF increasing the gene expression level; a repressing TF is similar in the sense that the expression level is decreased) (Table 1.8).

The first case is that only after all binding sites are bound, the cluster is able to alter the gene expression level. The cluster can only produce two different expression levels: zero or maximum (105, 106) (Table 1.8). Thus it prevents spurious transcription until the TF concentration is high enough such that all binding sites are occupied, and consequently reduces leaky gene expression and noise in mRNA levels (102, 107). In addition, it slows the initiation of gene transcription by requiring a longer time for all sites to become bound (102, 107).

The second case is that as long as one binding site is bound, the cluster is able to increase gene expression to the maximum level (Table 1.8). Thus it makes a promoter more sensitive to low concentrations of TFs and less sensitive to higher concentrations of TFs (102). In addition, it expedites gene expression by only requiring one site to become bound (102, 108).

The third case is that each binding site in a cluster independently and equally contributes to gene transcription, so that the gene expression level is proportional to the number of bound sites (102) (Table 1.8). *In vivo* this case does not always happen; different sites have different amounts of contribution, which is the fourth case (Table 1.8). For example, certain TFs have optimal distances from the TSS that maximize their interaction with the transcriptional machinery (109, 110). There may also be a periodic relation between the distance of a TFBS from the TSS and the level of transcription, possibly because the influence of TFs on gene expression is dependent on the nucleosome context (111).

1.5.1.1.2 Presence of Interaction Between Individual Sites

If direct or indirect interactions are present between individual binding site within a cluster, the gene expression level after all sites are bound exceeds the sum of the expression levels when each single site is bound (Table 1.8); that is, the clusters can

Table 1.8: Impacts of a homotypic TFBS cluster on the gene expression level

Cluster†		Interaction absent‡				Interaction present‡
1st site	2nd site	1st case	2nd case	3rd case	4th case	
×	×	0%	0%	0%	0%	0%
×	√	0%	100%	50%	25%	25%
√	×	0%	100%	50%	75%	25%
√	√	100%	100%	100%	100%	100%

† The cluster consists of two binding sites of the TF. × indicates that the site is not bound, and √ indicates that the site is bound.

‡ The percentage in each cell is the ratio of the expression level induced by the current binding situation of the cluster to the maximum expression level.

amplify the influence of individual binding sites on gene expression, through increased binding stability, funnel effects and facilitated diffusion mechanisms (102).

Highly degenerate TFBSs whose sequences differ much from the consensus sequence tend to occur in homotypic clusters (104). These adjacent weak binding sites, which individually may not be strong enough to bind TFs and activate transcription, can stabilize each other's binding by direct TF-TF dimerization and indirect nucleosome displacement (112, 113) (Figure 1.7). In addition, the weaker sites flanking a strong binding site in a cluster can direct the TF molecule to the strong site and extend the period of the molecule physically associating with the strong site, which is known as the funnel effect (114–116) (Figure 1.7).

TFs search for binding sites by a combination of three-dimensional diffusion in the nucleoplasm and one-dimensional random walk on the DNA, which is known as the facilitated diffusion mechanism (117–119) (Figure 1.7). The speedup in the search process results from the reduction of the dimensionality of the search process from three dimensions to one-dimension (102) (Figure 1.7). *In vitro* TFs associate with longer synthesized DNA fragments more rapidly compared to shorter DNA fragments that

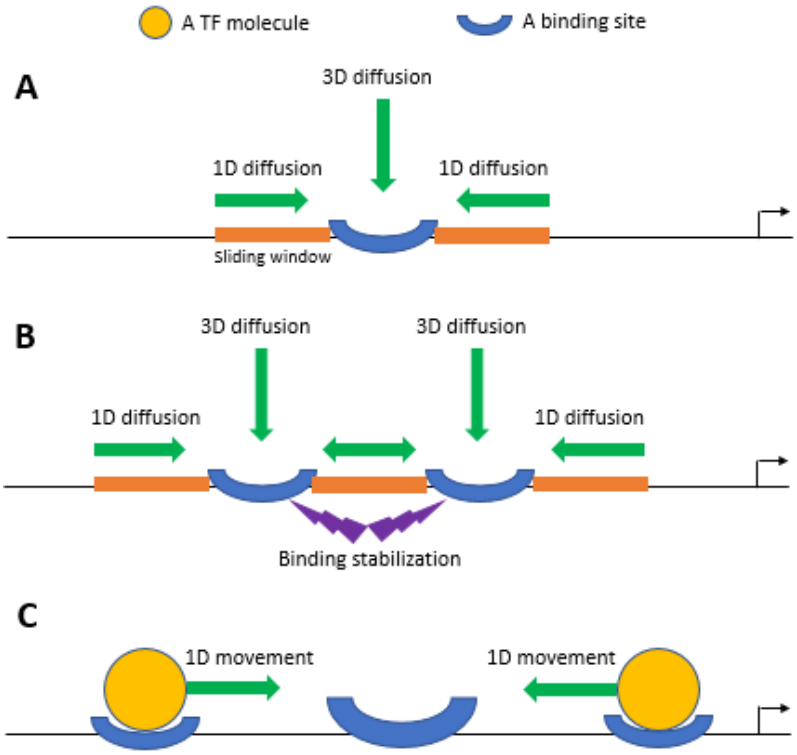

Figure 1.7: Cooperation between adjacent binding sites in a cluster. The size of a block arc positively correlates with the strength of the binding site. (A) If a TF molecule associates with DNA within the sliding window, it can perform a 1D random walk to find the binding site faster and with higher probability than the 3D diffusion. (B) Multiple adjacent weak binding sites within a cluster are able to stabilize each other's binding, and extend the sliding window to reduce TF search time. (C) TF molecules are directed to the central stronger site by the flanking weaker sites, which is known as the funnel effect.

contain the same binding site in the middle (120), which indirectly proves the existence of facilitated diffusion (102). This mechanism is called the antenna effect, which assumes that a longer DNA fragment increases the contribution of the one-dimensional random walk component to the TF search process (121).

1.5.1.2 Computational Algorithms Detecting Transcription Factor Binding Site Clusters

Previous studies have developed multiple algorithms to computationally detect TFBS clusters. These algorithms fall into two main categories, window-based and model-based (Table 1.9).

Table 1.9: Algorithms detecting TFBS clusters

Algorithm	Category	Method to detect TFBSs	Reference
SCORE	Window-based	Consensus sequences	(122)
CIS-ANALYST		PWMs	(123)
MSCAN			(124)
IDBC		iPWMs	(125)
Cister	Model-based	PWMs	(126)
Comet			(127)
MCAST			(128)
Cluster-Buster		Count matrices	(129)
Poisson distribution-based		Consensus sequences	(130)

The window-based algorithms, such as SCORE (122), CIS-ANALYST (123), MSCAN (124) and IDBC (125), rely on intuitive merging operations on initial windows created based on distances between binding sites. CIS-ANALYST and MSCAN applied PWMs to detect TFBSs, which allowed the binding site strengths to vary; whereas SCORE only used the predicted sites that exactly match the consensus sequences.

The model-based algorithms, including Cister (126), Comet (127), MCAST (128) and Cluster-Buster (129), apply probabilistic models (e.g. log likelihood ratios) to discriminate motif clusters from background DNA in a mathematically optimal way (129). All four methods are based on the Hidden Markov Model (HMM), which views locating regulatory regions in genomic DNA as a change-point problem, with the beginning of regulatory and non-regulatory regions corresponding to the change points (131). A statistical E-value is computed for each detected cluster to indicate its significance.

1.5.1.2.1 The Information Density-based Clustering Algorithm

The Information Density-Based Clustering (IDBC) algorithm will be used in Chapter 6 to effectively detect TFBS clusters by taking into account both the spatial organization (i.e. intersite distances) and information density (i.e. R_i values) of TFBSs (125). Its detailed steps include (Figure 1.8):

① For each binding site s, calculate the neighborhood information content (nic) as being the total of pairwise sums of the information content for s and each site lying within a radius d (number of bases) of s.

② For each site s with nic exceeding a threshold I, create an initial cluster in which s is the center and all binding sites within the radius d are included.

③ In the first phase of merging clusters, consider each pair of clusters with centers c_i and c_j.

- If either cluster contains the center of the other cluster, then merge the two clusters. The center of the new cluster is the stronger one between c_i and c_j. If they have the same R_i value, the center of the cluster containing more TFBSs is made the center of the new cluster, whereas the other center is relegated to being just a site.
- If c_i is included in the cluster with c_j as its center, but c_j is not included in the cluster with c_i as its center, compare the strengths of c_i and c_j.
    - If c_i is stronger than c_j, the overlapping TFBSs are put into the cluster with c_i as its center and removed from the cluster with c_j as its center.
    - If c_i is weaker than c_j, the overlapping sites are put into the cluster with c_j as its center and removed from the cluster with c_i as its center. The remaining sites in the cluster with c_j as its center form a cluster in which the strongest site c_k is selected as the new center.

This step is iterated until no c_i occurs in more than one cluster.

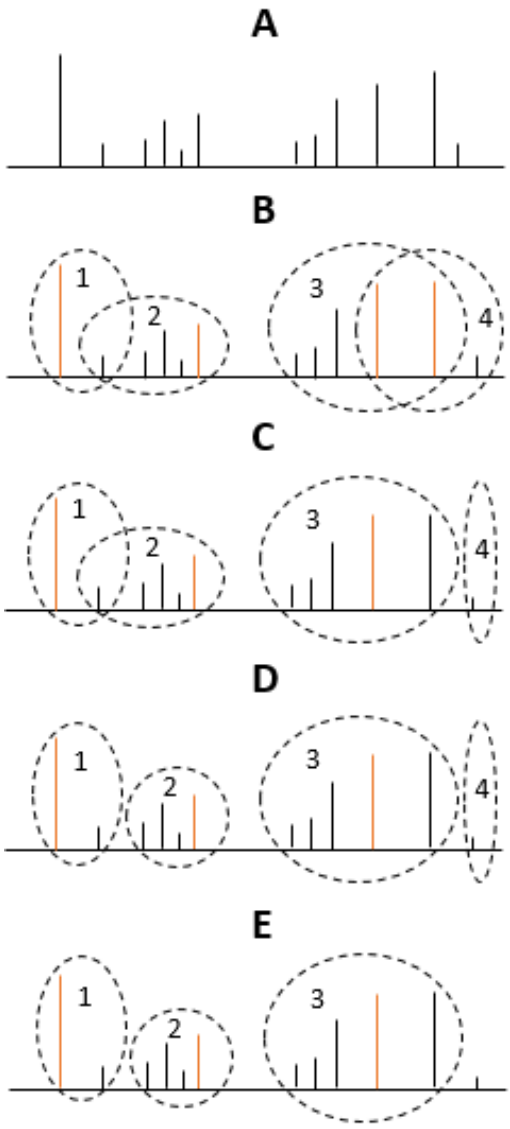

Figure 1.8: The IDBC algorithm. (A) Vertical bars indicate the locations of putative binding sites upstream of the TSS, and their heights are positively correlated with the strengths of the binding sites. (B) The initial four clusters are created from the first two steps of the algorithm. Clusters 1 and 2 share one binding site, and Clusters 3 and 4 contain each other's center. (C) Step 3 solves the overlap of Clusters 3 and 4 by putting the center of Cluster 4 into Cluster 3, since the center of Cluster 3 is of the same strength as that of Cluster 4 and Cluster 3 contains more sites. (D) Step 4 solves the overlapping site of Clusters 1 and 2 by putting it into Cluster 1. (E) In Step 5, since the single site in Cluster 4 is not strong enough to be a cluster, finally only three clusters remain.

④ In the second phase of merging, all binding sites that belong to more than one cluster are exclusively allocated to the cluster with the stronger center.

⑤ In the re-evaluation phase, a final check is performed to ensure that the information content of each cluster reaches the threshold *I* (as in Step 2) after the possible reallocation of sites in the preceding step. Clusters failing the check are dissolved into individual sites.

1.6 Prediction of Gene Expression Levels

As described above, the gene expression pattern is largely determined by the distribution and composition of TFBS clusters in the promoter, implying that TFBSs can be used to explain the variance in the expression level between genes. In fact, previous studies, based on either TF binding or chromatin structure data, predicted gene expression levels using a variety of machine learning classifiers and regression models. The TF binding profiles used to predict tissue-specific absolute gene expression levels were derived from either *in vivo* ChIP-seq peaks (132–134) or computationally detected binding sites and clusters (135). Both achieved similar accuracy (136).

Ouyang et al. (132) applied a principle component regression model based on features extracted from ChIP-seq peaks of 12 TFs in mouse embryonic stem cells, and found that these features explained 63.9% of the variance in the gene expression levels. The feature of each gene for each TF, termed as the TF association strength, was a weighted sum of signal values of all ChIP-seq peaks, where the weights are the distances of the peaks from the gene (132). Similarly, Cheng et al. (133) related the binding strengths of 40 TFs to the expression levels of 57 genes in the K562 cell line, and achieved a 67% predictive performance. The binding strength of each TF on each gene was computed by averaging the signal values of ChIP-seq peaks within a 100bp interval centered on the TSS (133).

Smith et al. (135) used the PWMs from the TRANSFAC database (85) to detect TFBSs and clusters, and applied the MARS algorithm (137) to construct a classification model for each of 56 human and mouse tissues to distinguish expressed from silent genes. The classifiers succeeded in 80% of the tissues with a smallest prediction error of 35%. The

TFBS clusters were formed by combining immediately adjacent two or three binding sites (135).

Chromatin structure data, including histone modifications (HMs) (e.g. H3K4me3, H3K9me3 and H3K27me3) and DNase I hypersensitivity, were also used to predict tissue-specific absolute gene expression levels (133, 134, 136, 138, 139). Similarly, the feature of a gene for a histone marker was defined as the sum of signal values of ChIP-seq peaks of the marker within a fixed-length interval around the TSS.

These studies found that chromatin structure is statistically redundant with TF binding in explaining tissue-specific mRNA transcript abundance at a genome-wide level, which was attributed to the heterogeneous distribution of HMs across chromatin domains (134). Combining these two types of data explained the largest fraction of variance in gene expression levels in multiple cell lines (133, 134), suggesting that either contributes unique information to gene expression that cannot be compensated for by the other.

These previous studies successfully related TF binding profiles and chromatin structure data to gene expression levels. However, they have several limitations:

1) Because signal values of ChIP-seq peaks are not strictly proportional to TFBS strengths contained in the peaks (2), representing TF binding strengths by ChIP-seq signals may not be appropriate.

2) The clustering algorithm used in (135) arbitrarily limits the number of binding sites contained in a module, which is inflexible. It does not consider the information densities of binding sites and clusters either.

3) When detecting TFBSs the PWMs in (135) are used to compute log likelihood ratio scores which are not Shannon information theory-based. These scores are not quantitatively related to the amount of dissipated heat energy by Equation 1.6, so that they are unable to quantify binding site strengths as accurately as R_i values.

4) The machine learning models in (135) are tissue-specific; each of the 56 tissues has a different classifier trained from genes expressed in the specific tissue. Thus each

classifier can only be applied to one single tissue, resulting in an excess of ungeneralizable models.

1.7 Prediction of Transcription Factor Target Genes

As described above, TF binding to target sites in the promoter results in the effective regulation of the gene expression level; however, not every binding event can alter the expression level. Cusanovich et al. (140) found that the number of genes directly bound by a TF significantly exceeds the number of genes whose expression levels significantly change upon knockdown of the TF. They performed siRNA-based knockdown experiments of 59 TFs in the GM19238 cell line, and measured the changes in the expression levels of 8,872 genes. They also indicated whether the promoters of these genes display evidence of binding to TFs by intersecting with ChIP-seq peaks in the GM12838 cell line, and observed that only a small subset of genes whose promoters overlap ChIP-seq peaks were differentially expressed (DE) after TF knockdown (140). Similarly, by perturbing expression of 10 TF genes with the CRISPR technique in K562 cells and performing single cell RNA sequencing, the regulatory effects of each TF on 22,046 genes were dissected with a regularized linear computational model (141), which accurately revealed DE targets and new functions of individual TFs.

Using the siRNA-generated knockdown data as the gold standard, correlation between TFBS counts and gene expression levels across 10 different cell lines were found to be more predictive of DE targets than setting a minimum threshold on TFBS counts (142). Machine learning classifiers have also been applied in a small number of gene instances to predict targets of a single TF using features extracted from *n*-grams derived from consensus binding sequences (143), or from TFBSs and homotypic binding site clusters (125).

These previous studies successfully used TF binding data to predict TF target genes. However, they also have a number of limitations.

1) In the correlation-based approach (142), TFBS counts were defined as the number of ChIP-seq peaks overlapping the promoter, though it was unknown how many binding sites were present in these peaks.

2) In the correlation-based approach (142), positives might not be direct targets in the TF regulatory cascade, as the promoters of these targets were not intersected with ChIP-seq peaks.

3) The machine learning approaches (125, 143) were applied on a small scale, rather than on the genome-wide set of target genes of multiple TFs identified from knockdown experiments.

1.8 References

1. Levine,M. and Tjian,R. (2003) Transcription regulation and animal diversity. *Nature*, **424**, 147–151.

2. Wang,J., Zhuang,J., Iyer,S., Lin,X., Whitfield,T.W., Greven,M.C., Pierce,B.G., Dong,X., Kundaje,A., Cheng,Y., *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, **22**, 1798–1812.

3. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.

4. Bi,C. and Rogan,P.K. (2004) Bipartite pattern discovery by entropy minimization-based multiple local alignment. *Nucleic Acids Res.*, **32**, 4979–4991.

5. Jolma,A., Yin,Y., Nitta,K.R., Dave,K., Popov,A., Taipale,M., Enge,M., Kivioja,T., Morgunova,E. and Taipale,J. (2015) DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, **527**, 384–388.

6. Parelho,V., Hadjur,S., Spivakov,M., Leleu,M., Sauer,S., Gregson,H.C., Jarmuz,A., Canzonetta,C., Webster,Z., Nesterova,T., *et al.* (2008) Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell*, **132**, 422–433.

7. Fleming,J.D., Pavesi,G., Benatti,P., Imbriano,C., Mantovani,R. and Struhl,K. (2013) NF-Y coassociates with FOS at promoters, enhancers, repetitive elements, and inactive chromatin regions, and is stereo-positioned with growth-controlling transcription factors. *Genome Res.*, **23**, 1195–1209.

8. Latchman,D.S. (1997) Transcription factors: an overview. *Int. J. Biochem. Cell Biol.*, **29**, 1305–1312.

9. Lescuyer,P., Martinez,P. and Lunardi,J. (2002) YY1 and Sp1 activate transcription of the human NDUFS8 gene encoding the mitochondrial complex I TYKY subunit. *Biochim. Biophys. Acta*, **1574**, 164–174.

10. Yao,Y.L., Yang,W.M. and Seto,E. (2001) Regulation of transcription factor YY1 by acetylation and deacetylation. *Mol. Cell. Biol.*, **21**, 5979–5991.

11. Rudnizky,S., Khamis,H., Malik,O., Squires,A.H., Meller,A., Melamed,P. and Kaplan,A. (2017) Single-molecule DNA unzipping reveals asymmetric modulation of a transcription factor by its binding site sequence and context. *Nucleic Acids Res.*, 10.1093/nar/gkx1252.

12. Schöne,S., Jurk,M., Helabad,M.B., Dror,I., Lebars,I., Kieffer,B., Imhof,P., Rohs,R., Vingron,M., Thomas-Chollier,M., *et al.* (2016) Sequences flanking the core-binding site modulate glucocorticoid receptor structure and activity. *Nat. Commun.*, **7**, 12621.

13. Dror,I., Golan,T., Levy,C., Rohs,R. and Mandel-Gutfreund,Y. (2015) A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res.*, **25**, 1268–1280.

14. Schneider,B., Božíková,P., Čech,P., Svozil,D. and Černý,J. (2017) A DNA Structural Alphabet Distinguishes Structural Features of DNA Bound to Regulatory Proteins and in the Nucleosome Core Particle. *Genes*, **8**.

15. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.*, **79**, 233–269.

16. Pastor,N., Pardo,L. and Weinstein,H. (1997) Does TATA matter? A structural exploration of the selectivity determinants in its complexes with TATA box-binding protein. *Biophys. J.*, **73**, 640–652.

17. Costello,J.F., Futscher,B.W., Kroes,R.A. and Pieper,R.O. (1994) Methylation-related chromatin structure is associated with exclusion of transcription factors from and suppressed expression of the O-6-methylguanine DNA methyltransferase gene in human glioma cell lines. *Mol. Cell. Biol.*, **14**, 6515–6521.

18. Benveniste,D., Sonntag,H.-J., Sanguinetti,G. and Sproul,D. (2014) Transcription factor binding predicts histone modifications in human cell lines. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 13367–13372.

19. Papavassiliou,A.G., Bohmann,K. and Bohmann,D. (1992) Determining the effect of inducible protein phosphorylation on the DNA-binding activity of transcription factors. *Anal. Biochem.*, **203**, 302–309.

20. Karin,M. and Chang,L. (2001) AP-1--glucocorticoid receptor crosstalk taken to a higher level. *J. Endocrinol.*, **169**, 447–451.

21. Barnes,P.J. and Karin,M. (1997) Nuclear factor-kappaB: a pivotal transcription factor in chronic inflammatory diseases. *N. Engl. J. Med.*, **336**, 1066–1071.

22. Liberman,A.C., Druker,J., Garcia,F.A., Holsboer,F. and Arzt,E. (2009) Intracellular molecular signaling. Basis for specificity to glucocorticoid anti-inflammatory actions. *Ann. N. Y. Acad. Sci.*, **1153**, 6–13.

23. Liberman,A.C., Refojo,D., Druker,J., Toscano,M., Rein,T., Holsboer,F. and Arzt,E. (2007) The activated glucocorticoid receptor inhibits the transcription factor T-bet by direct protein-protein interaction. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.*, **21**, 1177–1188.

24. Buckingham,J.C. (2006) Glucocorticoids: exemplars of multi-tasking. *Br. J. Pharmacol.*, **147 Suppl 1**, S258-268.

25. Harbers,M., Wahlström,G.M. and Vennström,B. (1996) Transactivation by the thyroid hormone receptor is dependent on the spacer sequence in hormone response elements containing directly repeated half-sites. *Nucleic Acids Res.*, **24**, 2252–2259.

26. Zaret,K.S. and Carroll,J.S. (2011) Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.*, **25**, 2227–2241.

27. Carroll,J.S., Liu,X.S., Brodsky,A.S., Li,W., Meyer,C.A., Szary,A.J., Eeckhoute,J., Shao,W., Hestermann,E.V., Geistlinger,T.R., *et al.* (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell*, **122**, 33–43.

28. Pihlajamaa,P., Sahu,B., Lyly,L., Aittomäki,V., Hautaniemi,S. and Jänne,O.A. (2014) Tissue-specific pioneer factors associate with androgen receptor cistromes and transcription programs. *EMBO J.*, **33**, 312–326.

29. Sekiya,T. and Zaret,K.S. (2007) Repression by Groucho/TLE/Grg proteins: genomic site recruitment generates compacted chromatin in vitro and impairs activator binding in vivo. *Mol. Cell*, **28**, 291–303.

30. Watts,J.A., Zhang,C., Klein-Szanto,A.J., Kormish,J.D., Fu,J., Zhang,M.Q. and Zaret,K.S. (2011) Study of FoxA pioneer factor at silent genes reveals Rfx-repressed enhancer at Cdx2 and a potential indicator of esophageal adenocarcinoma development. *PLoS Genet.*, **7**, e1002277.

31. Zaret,K.S. and Mango,S.E. (2016) Pioneer transcription factors, chromatin dynamics, and cell fate control. *Curr. Opin. Genet. Dev.*, **37**, 76–81.

32. West,A.G. and Sharrocks,A.D. (1999) MADS-box transcription factors adopt alternative mechanisms for bending DNA. *J. Mol. Biol.*, **286**, 1311–1323.

33. Horikoshi,M., Bertuccioli,C., Takada,R., Wang,J., Yamamoto,T. and Roeder,R.G. (1992) Transcription factor TFIID induces DNA bending upon binding to the TATA element. *Proc. Natl. Acad. Sci. U. S. A.*, **89**, 1060–1064.

34. Gietl,A., Holzmeister,P., Blombach,F., Schulz,S., von Voithenberg,L.V., Lamb,D.C., Werner,F., Tinnefeld,P. and Grohmann,D. (2014) Eukaryotic and archaeal TBP and TFB/TF(II)B follow different promoter DNA bending pathways. *Nucleic Acids Res.*, **42**, 6219–6231.

35. Rubio-Cosials,A., Battistini,F., Gansen,A., Cuppari,A., Bernadó,P., Orozco,M., Langowski,J., Tóth,K. and Solà,M. (2017) Protein Flexibility and Synergy of HMG Domains Underlie U-Turn Bending of DNA by TFAM in Solution. *Biophys. J.*, 10.1016/j.bpj.2017.11.3743.

36. Ngo,H.B., Kaiser,J.T. and Chan,D.C. (2011) The mitochondrial transcription and packaging factor Tfam imposes a U-turn on mitochondrial DNA. *Nat. Struct. Mol. Biol.*, **18**, 1290–1296.

37. Bazett-Jones,D.P., Leblanc,B., Herfort,M. and Moss,T. (1994) Short-range DNA looping by the Xenopus HMG-box transcription factor, xUBF. *Science*, **264**, 1134–1137.

38. Clark,D.J. and Leblanc,B.P. (2015) Analysis of DNA Supercoiling Induced by DNA-Protein Interactions. *Methods Mol. Biol. Clifton NJ*, **1334**, 161–172.

39. Koudelka,G.B. (1998) Recognition of DNA structure by 434 repressor. *Nucleic Acids Res.*, **26**, 669–675.

40. Shi,Y. and Berg,J.M. (1996) DNA unwinding induced by zinc finger protein binding. *Biochemistry (Mosc.)*, **35**, 3845–3848.

41. Pavletich,N.P. and Pabo,C.O. (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science*, **252**, 809–817.

42. Pavletich,N.P. and Pabo,C.O. (1993) Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. *Science*, **261**, 1701–1707.

43. Ozonov,E.A. and van Nimwegen,E. (2013) Nucleosome free regions in yeast promoters result from competitive binding of transcription factors that interact with chromatin modifiers. *PLoS Comput. Biol.*, **9**, e1003181.

44. Pellegrini,L., Tan,S. and Richmond,T.J. (1995) Structure of serum response factor core bound to DNA. *Nature*, **376**, 490–498.

45. West,A.G., Shore,P. and Sharrocks,A.D. (1997) DNA binding by MADS-box transcription factors: a molecular mechanism for differential DNA bending. *Mol. Cell. Biol.*, **17**, 2876–2887.

46. Stormo,G.D., Schneider,T.D., Gold,L. and Ehrenfeucht,A. (1982) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res.*, **10**, 2997–3011.

47. Shannon,C.E. (1948) A Mathematical Theory of Communication. *Bell Syst. Technol. J.*

48. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

49. Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415–431.

50. Rogan,P.K., Faux,B.M. and Schneider,T.D. (1998) Information analysis of human splice site mutations. *Hum. Mutat.*, **12**, 153–171.

51. Schneider,T.D. (1997) Information content of individual genetic sequences. *J. Theor. Biol.*, **189**, 427–441.

52. Basharin,G. (1959) On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables. *Theory Probab. Its Appl.*, **4**, 333–336.

53. Bi,C. and Rogan,P.K. (2005) Determining thresholds for binding site models using information theory. In. Salt Lake City, Utah.

54. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

55. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K., *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.

56. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G., *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.

57. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

58. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P., *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.

59. Li,Q., Brown,J.B., Huang,H. and Bickel,P.J. (2011) Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, **5**, 1752–1779.

60. Kharchenko,P.V., Tolstorukov,M.Y. and Park,P.J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, **26**, 1351–1359.

61. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol. ISMB Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

62. Bailey,T.L., Williams,N., Misleh,C. and Li,W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.

63. Machanick,P. and Bailey,T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinforma. Oxf. Engl.*, **27**, 1696–1697.

64. Setty,M. and Leslie,C.S. (2015) SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps. *PLoS Comput. Biol.*, **11**, e1004271.

65. Fratkin,E., Naughton,B.T., Brutlag,D.L. and Batzoglou,S. (2006) MotifCut: regulatory motifs finding with maximum density subgraphs. *Bioinforma. Oxf. Engl.*, **22**, e150-157.

66. Roth,F.P., Hughes,J.D., Estep,P.W. and Church,G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.

67. Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci. Publ. Protein Soc.*, **4**, 1618–1632.

68. Workman,C.T. and Stormo,G.D. (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.*

69. Frith,M.C., Hansen,U., Spouge,J.L. and Weng,Z. (2004) Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res.*, **32**, 189–200.

70. Frith,M.C., Saunders,N.F.W., Kobe,B. and Bailey,T.L. (2008) Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput. Biol.*, **4**, e1000071.

71. Thijs,G., Lescot,M., Marchal,K., Rombauts,S., De Moor,B., Rouzé,P. and Moreau,Y. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinforma. Oxf. Engl.*, **17**, 1113–1122.

72. Favorov,A.V., Gelfand,M.S., Gerasimova,A.V., Ravcheev,D.A., Mironov,A.A. and Makeev,V.J. (2005) A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics*, **21**, 2240–2245.

73. Liu,X.S., Brutlag,D.L. and Liu,J.S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.

74. Ettwiller,L., Paten,B., Ramialison,M., Birney,E. and Wittbrodt,J. (2007) Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nat. Methods*, **4**, 563–565.

75. Pavesi,G., Mereghetti,P., Mauri,G. and Pesole,G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.*, **32**, W199–W203.

76. Eskin,E. and Pevzner,P.A. (2002) Finding composite regulatory patterns in DNA sequences. *Bioinforma. Oxf. Engl.*, **18 Suppl 1**, S354-363.

77. Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–1659.

78. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.

79. Sinha,S. and Tompa,M. (2003) YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.*, **31**, 3586–3588.

80. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J., *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.

81. Kheradpour,P. and Kellis,M. (2014) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.*, **42**, 2976–2987.

82. Sandelin,A., Alkema,W., Engström,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.

83. Khan,A., Fornes,O., Stigliani,A., Gheorghe,M., Castro-Mondragon,J.A., van der Lee,R., Bessy,A., Chèneby,J., Kulkarni,S.R., Tan,G., *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**, D1284.

84. Wang,J., Zhuang,J., Iyer,S., Lin,X.-Y., Greven,M.C., Kim,B.-H., Moore,J., Pierce,B.G., Dong,X., Virgil,D., *et al.* (2013) Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res.*, **41**, D171-176.

85. Matys,V., Fricke,E., Geffers,R., Gössling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V., *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.

86. 1000 Genomes Project Consortium, Auton,A., Brooks,L.D., Durbin,R.M., Garrison,E.P., Kang,H.M., Korbel,J.O., Marchini,J.L., McCarthy,S., McVean,G.A., *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

87. Rogan,P.K., Svojanovsky,S. and Leeder,J.S. (2003) Information theory-based analysis of CYP2C19, CYP2D6 and CYP3A5 splicing mutations. *Pharmacogenetics*, **13**, 207–218.

88. Gadiraju,S., Vyhlidal,C.A., Leeder,J.S. and Rogan,P.K. (2003) Genome-wide prediction, display and refinement of binding sites with information theory-based models. *BMC Bioinformatics*, **4**, 38.

89. Antoniou,A.C. and Easton,D.F. (2006) Models of genetic susceptibility to breast cancer. *Oncogene*, **25**, 5898–5905.

90. Hollestelle,A., Wasielewski,M., Martens,J.W.M. and Schutte,M. (2010) Discovering moderate-risk breast cancer susceptibility genes. *Curr. Opin. Genet. Dev.*, **20**, 268–276.

91. Allikmets,R., Wasserman,W.W., Hutchinson,A., Smallwood,P., Nathans,J., Rogan,P.K., Schneider,T.D. and Dean,M. (1998) Organization of the ABCR gene: analysis of promoter and splice junction sequences. *Gene*, **215**, 111–122.

92. Twyman,R.M. (2004) SNP discovery and typing technologies for pharmacogenomics. *Curr. Top. Med. Chem.*, **4**, 1423–1431.

93. Hosseinpour,B., Bakhtiarizadeh,M.R., Khosravi,P. and Ebrahimie,E. (2013) Predicting distinct organization of transcription factor binding sites on the promoter regions: a new genome-based approach to expand human embryonic stem cell regulatory network. *Gene*, **531**, 212–219.

94. Shirley,B.C., Mucaki,E.J., Whitehead,T., Costea,P.I., Akan,P. and Rogan,P.K. (2013) Interpretation, Stratification and Evidence for Sequence Variants Affecting mRNA Splicing in Complete Human Genome Sequences. *Genomics Proteomics Bioinformatics*, **11**, 77.

95. Riva,A. (2012) Large-scale computational identification of regulatory SNPs with rSNP-MAPPER. *BMC Genomics*, **13**, S7.

96. Kumar,S., Ambrosini,G. and Bucher,P. (2017) SNP2TFBS – a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res.*, **45**, D139–D144.

97. Perera,D., Chacon,D., Thoms,J.A., Poulos,R.C., Shlien,A., Beck,D., Campbell,P.J., Pimanda,J.E. and Wong,J.W. (2014) OncoCis: annotation of cis-regulatory mutations in cancer. *Genome Biol.*, **15**, 485.

98. Andersen,M.C., Engström,P.G., Lithwick,S., Arenillas,D., Eriksson,P., Lenhard,B., Wasserman,W.W. and Odeberg,J. (2008) In silico detection of sequence variations modifying transcriptional regulation. *PLoS Comput. Biol.*, **4**, e5.

99. Zuo,C., Shin,S. and Keleş,S. (2015) atSNP: transcription factor binding affinity testing for regulatory SNP detection. *Bioinforma. Oxf. Engl.*, **31**, 3353–3355.

100. Thomas-Chollier,M., Hufton,A., Heinig,M., O'Keeffe,S., Masri,N.E., Roider,H.G., Manke,T. and Vingron,M. (2011) Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nat. Protoc.*, **6**, 1860–1869.

101. Coetzee,S.G., Coetzee,G.A. and Hazelett,D.J. (2015) motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinforma. Oxf. Engl.*, **31**, 3847–3849.

102. Ezer,D., Zabet,N.R. and Adryan,B. (2014) Homotypic clusters of transcription factor binding sites: A model system for understanding the physical mechanics of gene expression. *Comput. Struct. Biotechnol. J.*, **10**, 63–69.

103. Gotea,V., Visel,A., Westlund,J.M., Nobrega,M.A., Pennacchio,L.A. and Ovcharenko,I. (2010) Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res.*, **20**, 565–577.

104. Zhang,C., Xuan,Z., Otto,S., Hover,J.R., McCorkle,S.R., Mandel,G. and Zhang,M.Q. (2006) A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. *Nucleic Acids Res.*, **34**, 2238–2246.

105. Chu,D., Zabet,N.R. and Mitavskiy,B. (2009) Models of transcription factor binding: sensitivity of activation functions to model assumptions. *J. Theor. Biol.*, **257**, 419–429.

106. Bintu,L., Buchler,N.E., Garcia,H.G., Gerland,U., Hwa,T., Kondev,J. and Phillips,R. (2005) Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev.*, **15**, 116–124.

107. Zabet,N.R. and Chu,D.F. (2010) Computational limits to binary genes. *J. R. Soc. Interface*, **7**, 945–954.

108. Ezer,D., Zabet,N.R. and Adryan,B. (2014) Physical constraints determine the logic of bacterial promoter architectures. *Nucleic Acids Res.*, **42**, 4196–4207.

109. Cox,R.S., Surette,M.G. and Elowitz,M.B. (2007) Programming gene expression with combinatorial promoters. *Mol. Syst. Biol.*, **3**, 145.

110. Atkinson,T.J. and Halfon,M.S. (2014) Regulation of gene expression in the genomic context. *Comput. Struct. Biotechnol. J.*, **9**, e201401001.

111. Sharon,E., Kalma,Y., Sharp,A., Raveh-Sadka,T., Levo,M., Zeevi,D., Keren,L., Yakhini,Z., Weinberger,A. and Segal,E. (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.*, **30**, 521–530.

112. Wasson,T. and Hartemink,A.J. (2009) An ensemble model of competitive multi-factor binding of the genome. *Genome Res.*, **19**, 2101–2112.

113. Mirny,L.A. (2010) Nucleosome-mediated cooperativity between transcription factors. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 22534–22539.

114. Brackley,C.A., Cates,M.E. and Marenduzzo,D. (2012) Facilitated diffusion on mobile DNA: configurational traps and sequence heterogeneity. *Phys. Rev. Lett.*, **109**, 168103.

115. Weindl,J., Dawy,Z., Hanus,P., Zech,J. and Mueller,J.C. (2009) Modeling promoter search by E. coli RNA polymerase: one-dimensional diffusion in a sequence-dependent energy landscape. *J. Theor. Biol.*, **259**, 628–634.

116. Mirny,L., Slutsky,M., Wunderlich,Z., Tafvizi,A., Leith,J. and Kosmrlj,A. (2009) How a protein searches for its site on DNA: the mechanism of facilitated diffusion. *J. Phys. Math. Theor.*, **42**, 434013.

117. Zabet,N.R. and Adryan,B. (2012) Computational models for large-scale simulations of facilitated diffusion. *Mol. Biosyst.*, **8**, 2815–2827.

118. Berg,O.G., Winter,R.B. and von Hippel,P.H. (1981) Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry (Mosc.)*, **20**, 6929–6948.

119. von Hippel,P.H. and Berg,O.G. (1989) Facilitated target location in biological systems. *J. Biol. Chem.*, **264**, 675–678.

120. Kim,J.G., Takeda,Y., Matthews,B.W. and Anderson,W.F. (1987) Kinetic studies on Cro repressor-operator DNA interaction. *J. Mol. Biol.*, **196**, 149–158.

121. Shimamoto,N. (1999) One-dimensional diffusion of proteins along DNA. Its biological and chemical significance revealed by single-molecule measurements. *J. Biol. Chem.*, **274**, 15293–15296.

122. Rebeiz,M., Reeves,N.L. and Posakony,J.W. (2002) SCORE: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 9888–9893.

123. Berman,B.P., Nibu,Y., Pfeiffer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 757–762.

124. Alkema,W.B.L., Johansson,O., Lagergren,J. and Wasserman,W.W. (2004) MSCAN: identification of functional clusters of transcription factor binding sites. *Nucleic Acids Res.*, **32**, W195-198.

125. Dinakarpandian,D., Raheja,V., Mehta,S., Schuetz,E.G. and Rogan,P.K. (2005) Tandem machine learning for the identification of genes regulated by transcription factors. *BMC Bioinformatics*, **6**, 204.

126. Frith,M.C., Hansen,U. and Weng,Z. (2001) Detection of cis-element clusters in higher eukaryotic DNA. *Bioinforma. Oxf. Engl.*, **17**, 878–889.

127. Frith,M.C., Spouge,J.L., Hansen,U. and Weng,Z. (2002) Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res.*, **30**, 3214–3224.

128. Bailey,T.L. and Noble,W.S. (2003) Searching for statistically significant regulatory modules. *Bioinforma. Oxf. Engl.*, **19 Suppl 2**, ii16-25.

129. Frith,M.C., Li,M.C. and Weng,Z. (2003) Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.*, **31**, 3666–3668.

130. Wagner,A. (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinforma. Oxf. Engl.*, **15**, 776–784.

131. Crowley,E.M., Roeder,K. and Bina,M. (1997) A statistical model for locating regulatory regions in genomic DNA. *J. Mol. Biol.*, **268**, 8–14.

132. Ouyang,Z., Zhou,Q. and Wong,W.H. (2009) ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 21521–21526.

133. Cheng,C., Alexander,R., Min,R., Leng,J., Yip,K.Y., Rozowsky,J., Yan,K.-K., Dong,X., Djebali,S., Ruan,Y., *et al.* (2012) Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.*, **22**, 1658–1667.

134. Budden,D.M., Hurley,D.G., Cursons,J., Markham,J.F., Davis,M.J. and Crampin,E.J. (2014) Predicting expression: the complementary power of histone modification and transcription factor binding data. *Epigenetics Chromatin*, **7**, 36.

135. Smith,A.D., Sumazin,P., Xuan,Z. and Zhang,M.Q. (2006) DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 6275–6280.

136. McLeay,R.C., Lesluyes,T., Cuellar Partida,G. and Bailey,T.L. (2012) Genome-wide in silico prediction of gene expression. *Bioinforma. Oxf. Engl.*, **28**, 2789–2796.

137. Friedman,J.H. (1991) Multivariate Adaptive Regression Splines. *Ann. Stat.*, **19**, 1–67.

138. Karlić,R., Chung,H.-R., Lasserre,J., Vlahovicek,K. and Vingron,M. (2010) Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 2926–2931.

139. Dong,X., Greven,M.C., Kundaje,A., Djebali,S., Brown,J.B., Cheng,C., Gingeras,T.R., Gerstein,M., Guigó,R., Birney,E., *et al.* (2012) Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.*, **13**, R53.

140. Cusanovich,D.A., Pavlovic,B., Pritchard,J.K. and Gilad,Y. (2014) The functional consequences of variation in transcription factor binding. *PLoS Genet.*, **10**, e1004226.

141. Dixit,A., Parnas,O., Li,B., Chen,J., Fulco,C.P., Jerby-Arnon,L., Marjanovic,N.D., Dionne,D., Burks,T., Raychowdhury,R., *et al.* (2016) Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, **167**, 1853-1866.e17.

142. Banks,C.J., Joshi,A. and Michoel,T. (2016) Functional transcription factor target discovery via compendia of binding and expression profiles. *Sci. Rep.*, **6**, 20649.

143. Cui,S., Youn,E., Lee,J. and Maas,S.J. (2014) An improved systematic approach to predicting transcription factor target genes using support vector machine. *PloS One*, **9**, e94519.

Chapter 2

2 Thesis Overview

In this chapter, we will describe the objectives of this thesis, our methods to achieve these objectives, and the motivations for our methods.

2.1 Thesis Objectives

The objective of this thesis is to improve the current computational modelling of the transcriptional regulation of human genes. The ultimate goal of the transcriptional regulation is to mediate the accurate regulation of expression levels of TF target genes through the underlying physical interactions between TFs and their binding sites. Thus we dissected the transcriptional regulation into three levels; the lowest level is the individual binding site and its association with the TF protein, the intermediate level is the TFBS cluster and the cooperation between individual sites within it, and the highest level is the promoter and the cooperation between individual clusters within it.

Accordingly, we dissected the objective into five sub-objectives on these three levels. The two sub-objectives on the lowest level are to improve the derivation of TFBS motifs from ChIP-seq datasets, and to improve the detection of TFBS variants related to hereditary breast and ovarian cancer (HBOC) and the prediction of their effects using the iPWMs derived from these binding site motifs. The sub-objective on the intermediate level is to improve the modelling of the relationship between individual binding sites and the cluster in terms of composition and variation. The two sub-objectives on the highest level are to improve the modelling of the relationship between individual clusters and the promoter in terms of composition and variation, and to improve the prediction of gene expression patterns and TF target genes.

2.2 Our Methods

We proposed to conduct the following Study 1 to achieve the first sub-objective on the lowest level, Study 2 to achieve the second sub-objective on the lowest level, and Study 3 to achieve the remaining three sub-objectives.

1) In Chapter 3, when initially analyzing ChIP-seq datasets with Bipad, we noticed that it may return cofactor or noise motifs instead of primary motifs. This is attributable to the fact that Bipad was originally designed to analyze bacterial genomes, which contain few low-complexity sequences in contrast with human genomes. Thus we first aimed to improve the ability of Bipad to reveal primary motifs from human ChIP-seq datasets by recursively masking motifs found in previous iterations to discover additional conserved motifs from the same dataset, resulting in the Maskminent software. In the Maskminent motif discovery pipeline, we further sought to improve the detection of primary motifs and obtain the most accurate iPWMs by using the half-interval search to find the minimum threshold peak strength for inclusion of the maximum number of top peaks that can produce the primary motif. By applying the pipeline to ENCODE ChIP-seq peak datasets, we expected to obtain the iPWMs for sequence-specific TFs that enable prediction of TFBSs and mutation analyses on TFBS variants, and proposed to evaluate the accuracy of these iPWMs by detecting experimentally confirmed binding sites and explaining effects of experimentally characterized SNPs. We also expected to discover cofactor motifs, which support new TF-TF interactions and complexes.

2) In Chapters 4 and 5, we aimed to apply the iPWMs from Study 1 to detect TFBS variants in genes harboring known mutations that increase the risk of HBOC. We sought to prioritize these variants based on the extent of changes in the R_i values of binding sites caused by them.

3) In Chapter 6, we sought to apply the Bray-Curtis similarity metric to measure the similarity between genes in the tissue-wide expression profile. We further sought to develop a general machine learning framework that predicts genes with similar tissue-wide expression profiles to a given gene and predicts DE direct TF targets by combining information theory-based TF binding profiles with DNase I hypersensitive regions. We sought to derive TF binding features from clusters detected by the IDBC algorithm from iPWM-detected TFBSs that can effectively capture the spatial organization and informational composition of these clusters in the promoter. We also proposed to perform mutation analyses on promoters of target genes to investigate the downstream effects of

TFBS variants on information-dense clusters, the regulatory state and expression level of the gene.

2.3 Motivations for Our Methods

The methods proposed above aimed to overcome the limitations of the previous studies experimentally and computationally deriving TFBS motifs that were respectively described in Sections 1.3.1 and 1.3.2.3 of Chapter 1. Specifically, we sought to overcome Limitation 1 of both experimental and computational approaches by the fact that Bipad is capable of deriving bipartite binding site motifs from ChIP-seq datasets and generating bipartite iPWMs to accurately describe the binding behavior of dimeric TFs. We sought to avoid Limitations 2 and 3 of experimental approaches by the use of ChIP-seq datasets. We also sought to overcome Limitation 2 of computational approaches by analyzing the maximum number of top peaks that can produce the primary motif above the minimum threshold peak strength found by the half-interval search, Limitation 3 by the fact that Bipad is able to generate iPWMs, and Limitation 4 by the fact that Bipad does not rely on the background nucleotide composition and will always return the lowest-entropy motif.

The methods proposed above also aimed to overcome the limitations of the previous studies predicting gene expression levels and TF target genes that were respectively described in Sections 1.6 and 1.7 of Chapter 1. Specifically, we sought to overcome Limitation 1 of the prior studies predicting gene expression levels by the direct use of R_i values of binding sites, Limitation 2 by the use of the IDBC algorithm, Limitation 3 by the use of iPWMs, and Limitation 4 by the definition of the tissue-wide gene expression profile and application of the Bray-Curtis similarity measure. We also sought to overcome Limitation 1 of the prior studies predicting TF target genes by the direct use of iPWM-detected TFBSs, Limitation 2 by the fact that positives are direct targets whose promoters overlap tissue-specific ChIP-seq peaks, and Limitation 3 by the use of CRISPR- and siRNA-generated knockdown data.

Chapter 3

3 Discovery and Validation of Information Theory-based Transcription Factor and Cofactor Binding Site Motifs

The work presented in this chapter is reproduced from:

Lu,R., Mucaki,E.J., Rogan,P.K. (2017) Discovery and validation of information theory-based transcription factor and cofactor binding site motifs. *Nucleic Acids Research*, 45(5): e27

3.1 Introduction

Transcription factors (TFs) interact with regulatory elements in genes to mediate positive or negative regulation of tissue- and stage-specific expression (1, 2). TFs either directly bind to DNA by recognizing specific sequence motifs, or indirectly interact as partners (or cofactors) of sequence-specific TFs (3). Interactions between these two types of TFs, as well as between sequence-specific TFs, abound across the whole genome (3, 4). For instance, NF-Y extensively coassociates with FOS over all chromatin states, and CTCF extensively colocalizes with cohesins consisting of SMC1/SMC3 heterodimers and two non-SMC subunits RAD21 and SCC3 (5, 6). The genome-wide distributions of both types of bound TFs have been analyzed by sequence analysis of immunoprecipitated chromatin (ChIP-seq) (7). ChIP-seq can identify the repertoire of binding site sequences in a genome, and often pull down binding sites of coregulatory cofactors.

Sequence-specific TFs either recognize contiguous sequence motifs, or form homodimeric or heterodimeric structures that contact half sites separated by gaps that together comprise bipartite binding sites (8). Although generally the binding sequences of TFs are well conserved, significant variability at most positions of their binding motifs characterizes most TFs. Information theory-based position weight matrices (iPWMs) can quantitatively and accurately describe these base preferences. A contiguous iPWM is derived from a set of aligned binding sites using Shannon information theory and a uniform background nucleotide composition (9, 10). This approach may be more

appropriate for defining binding sites than Relative Entropy because the contacts between the TF and the nucleotides do not depend on the background genomic composition (10, 11). A bipartite iPWM consists of two contiguous, adjacent iPWMs, each corresponding to a half site, separated by a range of sequence gaps. The individual information content (R_i) of a TF-bound sequence, which represents the affinity of the TF-DNA interaction, is the dot product between the binary matrix of the sequence and an iPWM of the TF (10). The $R_{sequence}$ value of an iPWM is the mean of the R_i values of all the binding site sequences used to compute the iPWM, and represents the average binding affinity (12). Our laboratory previously developed the Bipad software to generate bipartite (and contiguous) iPWMs from ChIP-seq data (8).

TF binding motifs have been derived from both experimental evidence and computational approaches. Weirauch et al. (13) measured TF binding by octanucleotide microarrays to infer sequence specificity from overlapping bound sequences for >1,000 TFs encompassing 54 different DNA binding domain (DBD) classes. Jolma et al. (14) obtained 830 binding profiles representing 411 human and mouse TFs using high-throughput SELEX and ChIP sequencing. The oligonucleotide-based approach does not account for variable-length spacers in bipartite binding sites, and it may reconstruct potentially incorrect motifs that cannot be discriminated from correct binding site sequences. In addition, the set of octamers used in the DNA microarrays may not cover all possible binding site sequences (>8 nucleotides [nt]) recovered in the genome from ChIP-seq, and there is no way to discover potential binding sites from TF cofactors. Wang et al. (3) carried out *de novo* motif discovery for 119 human TFs from 457 ChIP-seq datasets using the MEME-ChIP software suite, and Kheradpour et al. (15) provided a systematic motif analysis for 427 ChIP-seq datasets of 123 human TFs using five motif discovery tools. However, these studies did not generate bipartite motifs with half sites separated by gaps varying in length; more importantly, the derived motifs were only based upon strongest ChIP-seq signal peaks (top 500 or 250 peaks), effectively eliminating thousands of intermediate or weak binding events and biasing the resulting iPWMs toward high-affinity, consensus-like binding sites. This is necessary, as the sequences contained in the weakest ChIP-seq peaks may contribute low-complexity,

likely non-functional sequences (i.e. noise) that can obfuscate the detection of true binding motifs. Extreme peak selection bias in the population of sites distorts the binding strengths estimated for individual sites (16).

We developed a motif discovery pipeline, Maskminent, by integrating recursive masking and thresholding the maximum number of ChIP-seq peaks into an entropy minimization framework. Bipad was modified to incorporate these features, and TF binding motifs were derived and validated for 765 ENCODE ChIP-seq datasets (1275 replicates) consisting of 207 human TFs. 93 primary and 23 cofactor binding motifs were successfully recovered and refined for 127 TFs. Reanalysis of the same data using the masking and thresholding techniques revealed many known and previously unreported TF cofactors; however, frequently our approach revealed cofactor motifs directly. These primary motifs were validated by comparing predicted with experimentally-detected true binding sites, explaining effects of characterized SNPs on binding site strengths, and through comparisons to an independent motif database.

3.2 Materials and Methods

3.2.1 ENCODE ChIP-seq datasets

The ENCODE Consortium conducted ChIP-seq assays for human TFs and generated initial peak datasets for each replicate of each assay using a uniform peak calling pipeline (7, 17). For some assays, these analyses produced optimal and conservative IDR-thresholded peaks after applying the IDR (Irreproducible Discovery Rate) framework to the initial datasets to improve consistency of motifs obtained from multiple biological replicates. In addition, Factorbook (3, 18) also reports motifs from refined datasets (limited to the top 500 peaks) generated by the SPP peak calling software (19).

We started with the IDR-thresholded peak datasets, because we found that these data are more likely to produce primary or cofactor motifs than the initial (i.e. unprocessed) datasets; they contain greater numbers of ChIP-seq peaks (and thus more binding sites) than the truncated SPP datasets. The initial, unfiltered datasets were examined if neither IDR-thresholded nor SPP datasets were available.

3.2.2 The Maskminent motif discovery pipeline

Initially, iPWMs from ChIP-seq reads were derived by entropy minimization with Bipad (Appendix B.1). However, we noted that these iPWMs sometimes exhibited cofactor or noise motifs, rather than the expected primary motifs. In order to improve detection of primary motifs, the Maskminent software, which implements a generalization of the objective function used in Bipad, enables new motif discovery by recursively masking sequences detected by previous analyses of a ChIP-seq dataset while defining thresholds for inclusion of the maximum number of top peaks to eliminate peaks with lower signal intensities whose inclusion can result in emergence of noise over primary or cofactor motifs (Appendix B.1). Multiple ChIP-seq datasets from distinct cell lines for the same TF, if available, were examined for enriched sequence motifs to assess whether this approach was reproducible, and discover tissue-specific sequence preferences between these sources.

This masking technique, which contrasts with the likelihood approach used by MEME (20), provides a means of discovering additional conserved motifs adjacent to primary TF binding sites within the same datasets. The sequences detected by motifs found in previous iterations are masked and the next lowest entropy motif is derived. The coordinates of all the predicted binding sites in a dataset scanned with prior iPWMs are recorded and skipped in the subsequent reanalysis. The specified parameters include the length of the motif, number of Monte Carlo cycles used in entropy minimization, a motif masking file for recursion, and for bipartite binding sites, the lengths of the left and right motifs and the gap length range between the half sites (Appendix B.1). Once a motif is generated, another program, Scan, is used to detect binding sites in a DNA sequence and determine their respective information contents, or binding strengths.

To eliminate noisy patterns that suppress the expected TF binding motifs due to ChIP-seq peaks with low signal strengths (i.e. read counts), the dataset is truncated based on signal strengths as follows (Figure 3.1). First, all the peaks are ranked in the descending order of strengths, and the top 200 peaks are selected. If the iPWM derived from the top 200 peaks exhibits the primary/cofactor motif, then the minimum threshold peak strength is contained within the range from the strength of the 200th peak (i.e. the initial value of G)

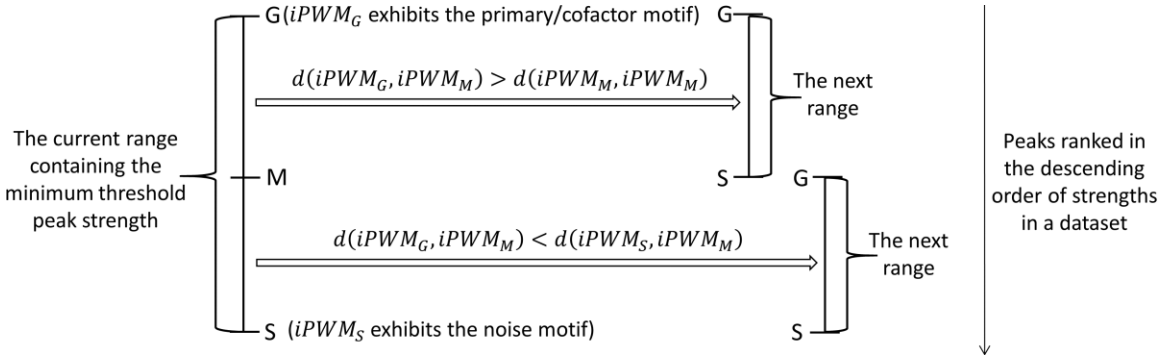

Figure 3.1: One iteration of the half-interval search used to refine the threshold peak strength. All peaks in the dataset are sorted in the descending order of signal strengths. $\boldsymbol{S}$ is the smaller bound of the current range containing the minimum threshold that can generate the primary/cofactor motif, and $\boldsymbol{G}$ is the greater bound (i.e. the current threshold). $\boldsymbol{G}$ and $\boldsymbol{S}$ are respectively initialized to the strength of the 200th peak and the strength of the last peak. $\boldsymbol{M}$ is the strength of the peak at the mean (rounding to the nearest multiple of 500) of the number of top peaks above $\boldsymbol{G}$ and the number of top peaks above $\boldsymbol{S}$. $\boldsymbol{iPWM_G}$, $\boldsymbol{iPWM_S}$, $\boldsymbol{iPWM_M}$ are respectively the iPWMs derived from the top peaks above $\boldsymbol{G}$, $\boldsymbol{S}$, $\boldsymbol{M}$. $\boldsymbol{d(iPWM_G, iPWM_M)}$ is the Euclidean distance between $\boldsymbol{iPWM_G}$ and $\boldsymbol{iPWM_M}$, and $\boldsymbol{d(iPWM_S, iPWM_M)}$ is the Euclidean distance between $\boldsymbol{iPWM_S}$ and $\boldsymbol{iPWM_M}$. If $\boldsymbol{d(iPWM_G, iPWM_M)}$ is greater than $\boldsymbol{d(iPWM_S, iPWM_M)}$, $\boldsymbol{iPWM_M}$ exhibits the noise motif and the minimum threshold is contained in the subrange from $\boldsymbol{G}$ to $\boldsymbol{M}$; if $\boldsymbol{d(iPWM_G, iPWM_M)}$ is smaller than $\boldsymbol{d(iPWM_S, iPWM_M)}$, $\boldsymbol{iPWM_M}$ exhibits the primary/cofactor motif and the minimum threshold is contained in the subrange from $\boldsymbol{M}$ to $\boldsymbol{S}$. When the number of peaks contained in the range does not exceed 500, this half-interval search is stopped. The approximately minimum threshold that is returned is $\boldsymbol{G}$ of the final range.

to the peak with the weakest signal (i.e. the initial value of S). A half-interval search iterated over sets of progressively weaker peaks narrows this range until the number of peaks contained in the range is ≤500. The value of G is the threshold peak signal strength above which the top peaks can still produce the primary/cofactor motif. The minimum threshold obtained for G (i.e. the final value of G) defines the approximate peak set containing the maximum number of top peaks that can produce the primary/cofactor motif.

3.2.3 Binding site motif validation

The methods used to evaluate the accuracy of our iPWMs include:

1) To detect experimentally proven binding sites in known target genes, derived iPWMs were used to evaluate the R_i value of each site;

2) To predict changes in binding site strength, characterized variants were evaluated with the corresponding iPWMs. The predicted changes were compared with experimentally supported effects on TF binding or gene expression;

3) The iPWMs were compared with the corresponding annotated motifs in the CIS-BP database (13) based on their normalized Euclidean distances;

4) To distinguish true binding motifs from noise motifs, we delineated the relationship between R_i values of binding sites discovered by the iPWM and their corresponding binding energy (i.e. higher R_i values have lower binding energies) (Appendix B.1). Primary/cofactor motifs are expected to demonstrate this relationship, whereas noise motifs are not; that is, for primary/cofactor motifs, the linear regression fit between R_i values and binding energy are expected to have slopes well below 0 which is the expected slope for noise motifs. After applying F-tests to evaluate this relationship, F values for the two categories of motifs were compared using a Mann-Whitney U test.

3.3 Results

The derived iPWMs displayed primary motifs for 93 TFs (Appendix B.2), as well as 23 cofactor motifs for 127 primary TFs (Appendix B.3). We also describe 6 high-confidence

novel motifs that have not been previously annotated in these ChIP-seq data (Appendix B.4).

The initial iPWMs directly exhibited primary motifs for 76 TFs and 18 cofactor motifs for 107 primary TFs. Thresholding the datasets revealed 31 primary motifs and 14 cofactors for 38 primary TFs. We used the masking technique to discover an additional 4 primary motifs; 7 cofactor motifs were also found in 21 datasets (Appendix B.2&B.3).

For each TF ChIP-seq dataset with a derived primary motif (n=367), we determined the false positive detection rate from the null R_i distribution, which is approximately Gaussian (12). The iPWM was used to scan for binding sites in a random 10,000 nucleotide sequence that conserved the mono- and dinucleotide composition as the dataset (Appendix B.2). The means of all null distributions range from -97.5 to -12.3 bits with standard deviations from 6.9 to 22.5 bits. The probabilities of observing a potentially functional binding site, i.e. with R_i>0, in these sequences range from 1.2E-7 to 0.06.

Similarly, the independence of contributions of each position in a binding site to the overall information content was analyzed for one iPWM of each primary motif. The total mutual information, which measures the interdependence between individual positions in the same binding site, was determined by summing the pairwise mutual information at each position (Appendix B.2). Then, the percentage of the total mutual information relative to the average information, $R_{sequence}$, was determined. For 83 TFs (~89.2%), <10% of the information present in the iPWM is dependent, and for 62 TFs (~66.7%), <5% is dependent. Neglecting the interactions between positions introduces a minimal error into the calculation of R_i values of binding sites, and would be expected to have little impact on assessment of the mutations in these sequences.

3.3.1 Primary binding motifs

3.3.1.1 Contiguous iPWMs

Correct iPWMs were successfully derived for 65 TFs with contiguous binding motifs, which are concordant with published descriptions of these motifs (3). All of these motifs can be characterized as degenerate and do not correspond to published consensus

sequences. Consensus sequences miss TF binding sites of weak or intermediate strength (16). We determined the frequencies of such sequences appearing on a genome scale for 10 TFs by counting the peaks containing these sequences in their respective datasets (Figure 3.2 - panel A). Surprisingly, only 0.015%-7.3% of all peaks contain binding sites with these sequences, demonstrating that these sites are extremely rare in ChIP-seq datasets. Thus, intermediate and low-affinity TF-DNA interactions are the most prevalent *in vivo* and are able to regulate gene expression (21).

3.3.1.2 Bipartite iPWMs

For 19 TFs, bipartite iPWMs were successfully derived, and were in agreement with previously reported motifs. The following examples illustrate key insights that can be taken from bipartite modeling:

1) El Marzouk et al. (22) demonstrated that ESR1 is able to recognize binding sites with half sites separated by nucleotide spacer lengths from 0 - 4nt, in which sites containing a 3nt spacer are most common and have the highest binding affinities. We allowed the spacer length to vary from 0 to 5nt in bipartite iPWMs derived from the T47D cell line data. The resultant iPWMs show the documented predominant sequences and are palindromic. The bipartite iPWM exceeds the average information content of the corresponding contiguous iPWM prepared from the same dataset, and the dominant gap between half sites is 3nt (Figure 3.2 - panel B). Nevertheless, 333 binding sites (~9%) in this iPWM exhibit a 5nt spacer, implying that ESR1 may be capable of binding to sites that were not previously detected. The symmetry between the half sites exhibited by the bipartite iPWMs suggests that dimeric ESR1 may bind a narrow range of sequences with similar half site affinities.

2) The palindromic predominant sequence of the AP2 family is 5'-GCCN3GGC-3', and other binding sequences confirmed in an *in vitro* binding-site selection assay include 5'-GCCN4GGC-3' and 5'-GCCN3/4GGG-3'. Another binding site 5'-CCCCAGGC-3' was also found in the SV40 enhancer (23). The spacer lengths in the bipartite iPWMs for AP2A and AP2C range from 2 – 4nt, which is representative of the genome-wide pool of true binding sites (Figure 3.2 – panel B). We also noted that the two outermost positions

are the most variable, and that adenine (instead of the consensus guanine) can also appear at the first position of the right half site. These bipartite iPWMs exhibit similar conservation levels across all the individual positions, suggesting that these binding sites of the two AP2 members may exhibit similar degrees of binding affinity, though iPWMs can recognize different sequences.

3) The predominant spacer length separating half sites recognized by STAT1 is 3nt; however, previous reports describe sites with a 2nt gap, but not those separated by 4nt (24). However, the STAT1 bipartite iPWM is based on 1709 binding sites (~18%) with a 4nt spacer, with most half sites separated by 2 or 3 nt (Figure 3.2 – panel B). The left- and rightmost nucleotides are nearly invariant, whereas the inner 2 nucleotide contacts in each half site are variable.

4) NFE2 and BACH1 heterodimerize with the MAF family (MAFF, MAFG and MAFK), and recognize two types of bipartite palindromic motifs, defined by the predominant binding sites TGCTGA(C)TCAGCA and TGCTGA(CG)TCAGCA (25). The previously reported binding motifs (3) are contiguous, and do not account for the dimeric interaction that gives rise to this bipartite binding pattern. The bipartite iPWMs indicate that the inner 6 positions surrounding the dominant 1nt spacer exhibit higher information contents than the outer 6 positions (Figure 3.2 – panel B).

3.3.1.3 Comparing iPWMs for the same TF in distinct cell lines

Cell-type-specific differences between iPWMs of the same TF were evident for certain contiguous and bipartite motifs. For instance, among the three contiguous iPWMs of ESR1 derived from the ECC1 steroid-responsive endometrial cell line, conservation levels in the respective half sites are asymmetric, whereas the average information of these half sites are much more symmetric in iPWMs derived from T47D, a breast tumor cell line (Figure 3.3 – panel A). For the TFs MAFF and MAFK, the discrepancy between the bipartite iPWMs from K562 and HepG2 cells is evident: the outer 6 positions show a greater degree of conservation than the internal 6 positions in HepG2, but in K562 the opposite trend is illustrated (Figure 3.3 – panel A). The MAFK iPWM derived from ChIP-seq data of IMR90 cells resembles the HepG2 iPWMs, whereas the iPWMs from

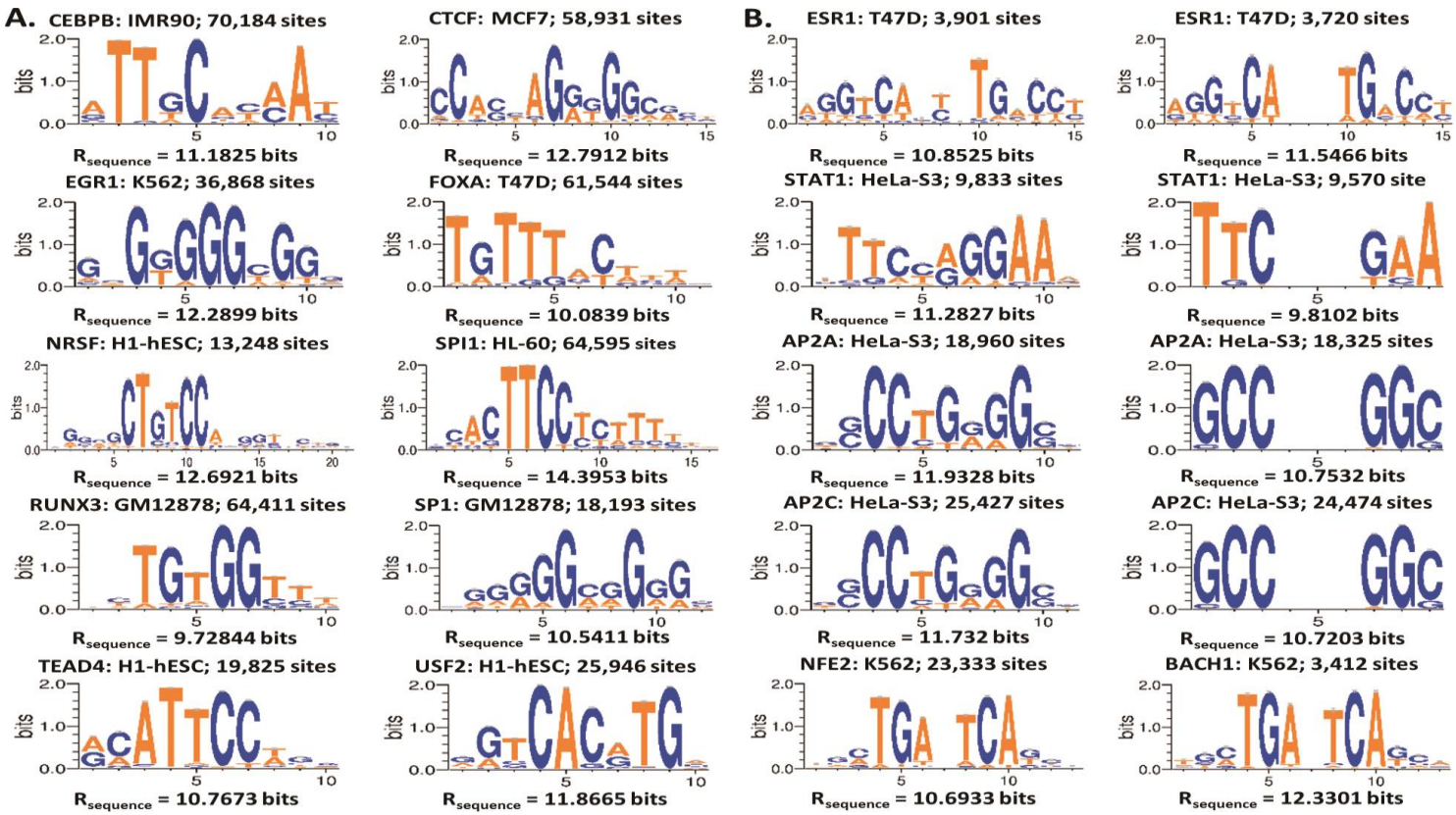

Figure 3.2: Sequence logos of contiguous (A) and bipartite (B) iPWMs. The TF name, and the cell line from which the iPWM was derived, and the number of binding sites that the iPWM is based upon are displayed. In (B), each of the first four rows includes a contiguous (left) iPWM and a bipartite (right) iPWM of one TF from the same dataset. The last row includes the bipartite iPWMs of NFE2 and BACH1. The bipartite search patterns, which are denoted by l<a,b>r (l and r are the lengths of the left and right half sites respectively, a and b are the minimum and maximum spacer lengths respectively), are 6<0,5>6, 3<2,4>3, 3<2,4>3, 3<2,4>3, 6<1,2>6 and 6<1,2>6 from top to bottom, respectively.

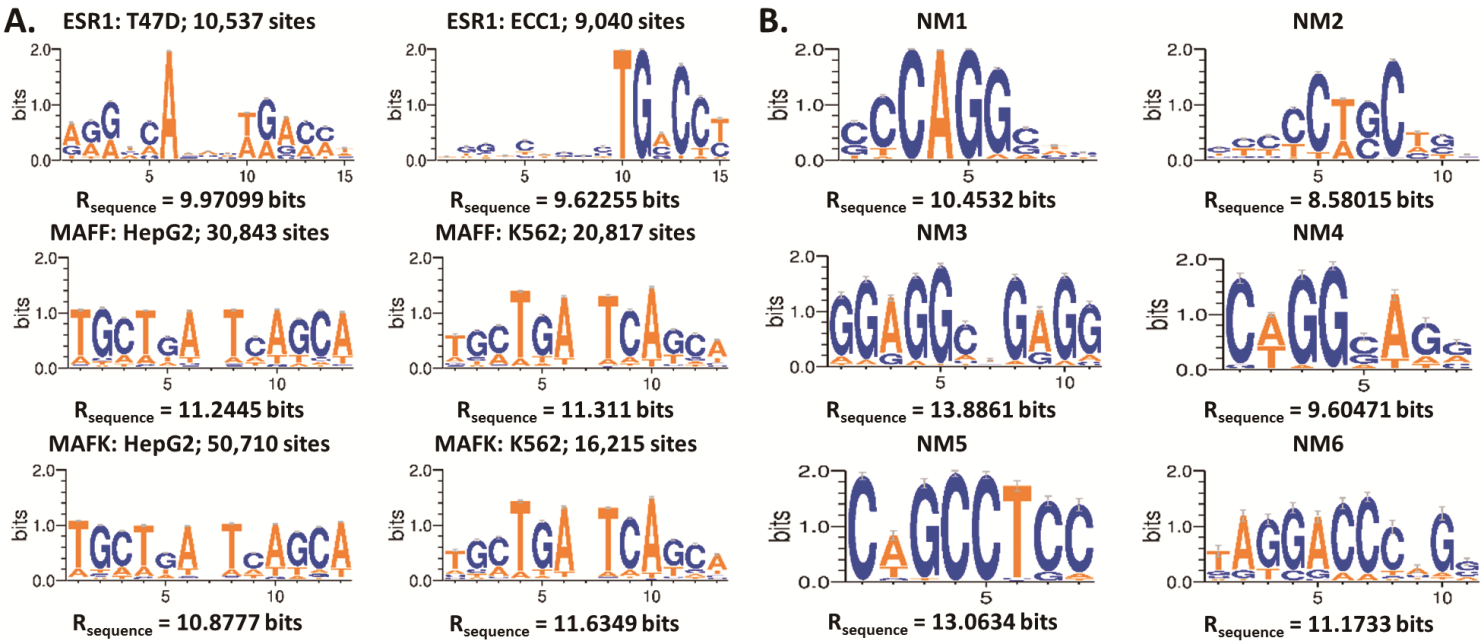

Figure 3.3: Comparison between iPWMs from different cell lines and novel motifs. (A) Each row includes sequence logos of two iPWMs of the same TF from two different cell lines. The bipartite iPWMs for MAFF and MAFK used the search pattern 6<1,2>6. (B) The high-confidence novel motifs ("NM1" – "NM6"). The logos of the NM1, NM2 and NM3 motifs come from the datasets of BAF155, NANOG and ESRRA, respectively**.**

HeLa-S3 and H1-hESC datasets resemble the K562 iPWMs. The compositions of binding sites (i.e. different target genes for the same TF in different tissues) account for these differences because TFs can display distinct cell-type-specific DNA sequence preferences (26). Consistent iPWMs between replicate datasets makes it unlikely that the skewed base conservation between ChIP-seq datasets for the same TF in different cell lines arises from sampling differences; however, this possibility cannot be excluded.

3.3.2 Cofactor binding motifs

Discovery of the binding motif of a cofactor in the same ChIP-seq dataset for a primary TF implies that the two TFs transcriptionally co-regulate this set of common target genes. This could be accomplished either by formation of a physical complex on the promoter, or by synergistic or antagonistic cis-regulatory effects. De novo motif discovery from ChIP-seq datasets provides an effective approach for confirming or predicting statistically significant TF interactions on a genome-wide scale; by contrast, the abundant, existing literature overwhelmingly documents gene-by-gene evidence about such interactions which constrains arguments supporting their generalizability. Figure 3.4 illustrates TF-cofactor interactions revealed by the Maskminent pipeline.

3.3.2.1 Confirmation of known cofactors

The derived iPWMs confirmed genome-wide interactions between 22 cofactors and 102 primary TFs (Table 3.1), which were supported by the previous studies (3, 5, 6, 15, 27-93) . For example, the interaction between SP1 and multiple members of the ETS and AP1 families has been well characterized (94–99). ELK1 and SRF can recruit each other to form a ternary complex on CArG-ETS elements (100). TEAD-AP1 cooperation with SRC coactivators drives downstream gene transcription to regulate cancer cell migration and invasion (101), and STAT1, STAT2 and IRF9 form a heterotrimer that regulates transcription of genes containing IFN-stimulated response elements (ISREs) (102). Consistent with previous reports (15), the existence of a YY1-THAP1 complex is predicted from co-segregation of their binding motifs in the K562 dataset of THAP1. Similarly, we predict that the SOX2-OCT4 complex colocalizes with BCL11A, similar to Wang et al (3). A DNA-binding complex consisting of GATA1, TAL1, E2A, LMO2 and

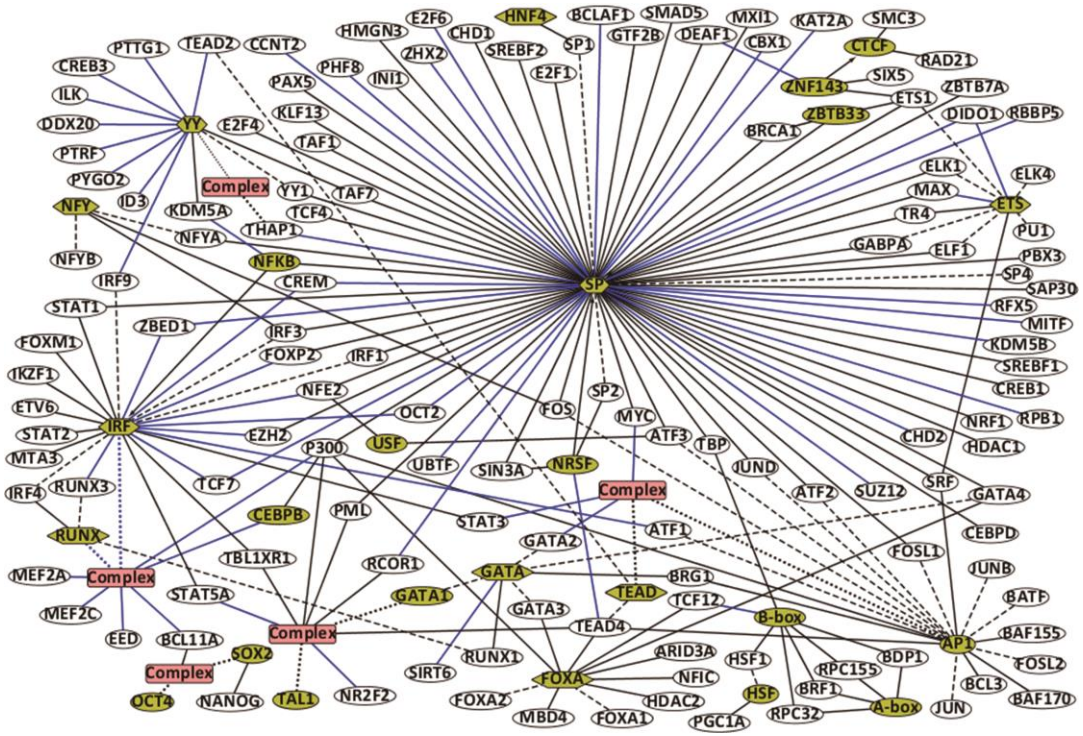

Figure 3.4: Network graph of TF-cofactor interactions revealed by the Maskminent pipeline. A yellow ellipse denotes a cofactor and a white ellipse denotes a primary TF. A hexagon denotes a TF family with dash lines connecting its members. For a TF family only members for which ENCODE provides peak datasets are shown. A red rectangle denotes a known or predicted TF complex with black or blue dotted lines indicating its components, respectively. An undirected line denotes the interaction between a primary TF and a cofactor which may be a complex or a TF family. A directed line links two cofactors, denoting that in a dataset of the starting TF, the ending TF was discovered as a cofactor. Black lines denote known interactions and blue lines denote the newly discovered interactions.

Table 3.1: Cofactors revealed by iPWMs and their corresponding primary TFs

Cofactors	Primary TFs*	
	Sequence-specific	**Non-sequence-specific**
AP1	GATA2, MYC, SRF, STAT3, TEAD4	BAF155, BAF170, BCL3, BRG1, P300
CEBPB		P300
CTCF	ZNF143	RAD21, SMC3
ETS family	MAX, SRF[1], TR4	DIDO1[2]
GATA family	RUNX1[2]	BRG1[2], SIRT6[2]
GATA1-TAL1	NR2F2[2], STAT5A[2], TAL1[2], TEAD4[2]	P300[2], PML[2], RCOR1[2], TBL1XR1[2]
FOXA family	ARID3A[3], GATA3, GATA4[3], NFIC[3], TCF12[3], TEAD4[3]	HDAC2[3], MBD4[3], P300
HNF4 family	SP1[3]	
HSF family		PGC1A[3]
IRF family	ATF1[2], BCL11A[1], CEBPB[1], CREM[1], ETV6[1], FOXM1[1], FOXP2, IKZF1[1], MEF2A[1], MEF2C[1], NFE2[1], NFKB[1], OCT2[1], RUNX3[1], STAT1[2], STAT2[2], STAT3[1], STAT5A[1], TCF7[1], ZBED1[1]	EED[1], EZH2[1], MTA3[1], P300[1], TBL1XR1[1]
NFKB		KDM5A[4]
NFY	FOS, IRF3	
NRSF	SP2[3], TEAD4	SIN3A[4]

RUNX family	BCL11A[1], CEBPB[1], IRF4[1], MEF2A[1], MEF2C[1]	EED[1], P300[1]
SP family	ATF2[4], ATF3, CEBPD[3], CREB1, CREM[1], DEAF1[2], E2F1, E2F4, E2F6, ELF1, ELK1, ETS1, FOS, FOSL1[4], FOXP2, GABPA, GATA4[3], IRF1[2], IRF3, JUND, KLF13[2], MAX, MITF[2], MXI1, MYC, NFE2[1], NFKB[1], NFYA, NRF1, NRSF[3], OCT2[1], PAX5[1], PBX3, RFX5, SMAD5, SREBF1[3], SREBF2[3], SRF, STAT1[1], SUZ12, TBP, TCF4, TCF7[2], THAP1[2], TR4, UBTF[2], YY1, ZBED1[2], ZBTB33, ZBTB7A[2], ZHX2[3]	BCLAF1, BRCA1, CBX1[3], CCNT2[2], CHD1, CHD2, DIDO1[2], EZH2, GTF2B[2], HDAC1[2], HMGN3[2], INI1, KAT2A, KDM5B[2], P300[4], PHF8[2], PML, RBBP5, RCOR1[3], RPB1, SAP30[2], SIN3A, TAF1, TAF7
SOX2	NANOG[4]	
SOX2-OCT4	BCL11A[4], OCT4[4]	
TEAD family	GATA2, MYC, STAT3	
TFIIIC	HSF1[3], TBP, TCF12	BDP1, BRF1, RPC155, RPC32
YY family	CREB3[2], IRF9[2], PTTG1[2], TEAD2[2], THAP1[2]	DDX20[2], ID3[2], ILK[2], KDM5A[4], PTRF[2], PYGO2[2], TAF7[2]
USF	ATF3, NFE2[1]	
ZBTB33	ETS1[1]	BRCA1
ZNF143	ETS1, DEAF1[2]	SIX5

* The underlined or normal font denotes known or newly discovered interactions between cofactors and primary TFs, respectively.

[1,2,3,4] The cofactor was revealed in the GM12878-related, K562, HepG2 or H1-HESC cell lines, respectively. Otherwise the cofactor appeared in other or multiple cell lines.

LDB1 is present in the erythroid cell lineage (103). Based on the proximity and coprecipitation of these binding sequences, we and others (3, 104) find that this complex, in which GATA1 and TAL1 contact DNA, coordinately binds with TEAD4 and other non-DNA binding proteins (P300, PML, RCOR1 and TBL1XR1). The GATA1-TAL1 and SOX2-OCT4 complexes emerged from the datasets of TAL1 and OCT4 as primary motifs, respectively, which implies the formation of the two complexes being necessary for binding of TAL1 and OCT4.

3.3.2.2 Discovery of novel cofactors

Maskminent revealed a number of previously unrecognized cofactor motifs (n=10) for 46 primary TFs (Table 3.1), which supports novel TF cobinding and interactions. This includes possible associations between the IRF and RUNX families, and their further cooperation with BCL11A, MEF2A, MEF2C, CEBPB, EED and P300 in GM12878 cells (Table 3.1; Figure 3.4). Similarly, the TEAD-AP1 complex is predicted to recruit MYC, STAT3 and GATA2 in multiple cell lines. The finding that NR2F2 and STAT5A motifs are in close proximity to sequences recognized by the GATA1-TAL1 complex suggests these factors may coordinately regulate target genes. Many cofactors were also discovered among datasets of non-sequence-specific primary TFs, which is consistent with the possibility that these primary TFs are recruited to gene promoters through their association with DNA-binding cofactors (Table 3.1).

3.3.2.3 Cofactor binding sites

To validate the predicted cobinding between cofactors and primary TFs, we determined the intersite distance distributions by scanning the individual ChIP-seq intervals with the derived iPWMs for each (Figure 3.5; Appendix B.5). A minimum information threshold was applied to the R_i values of predicted binding sites in order to remove the relatively large number of weak binding sites that are likely to be low-complexity sequences (e.g. $R_{sequence}$ [or 0.5 * $R_{sequence}$, if too many cofactor binding sites were eliminated at the higher threshold]). The SOX2-OCT4 complex was used as a primary negative control, as it is primarily expressed in the H1-hESC cell line and is unlikely to be a cofactor for primary TFs in other cell lines. A large percentage of peaks have short intersite distances

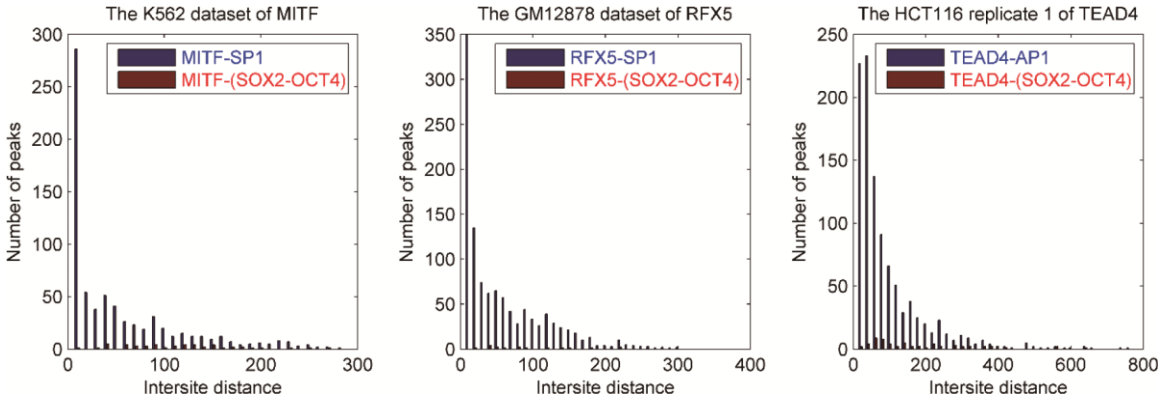

Figure 3.5: Distributions of intersite distances between primary TFs and discovered cofactors versus negative controls. The minimum threshold on information contents of predicted binding sites is $\boldsymbol{R_{sequence}}$. Each graph illustrates a much higher frequency of short (< 20nt) intersite distances between primary TFs and cofactors (blue) compared to the negative control (SOX2-OCT4; red).

between the primary TF and the corresponding cofactor binding sites (e.g. <20nt), whereas there is no such a trend for the negative control sequences and the primary TF. The same difference is observed between the distribution for the documented TEAD4-AP1 pair and for the negative control. Consistent with previous reports (4), the binding sites of cofactors and primary TFs in peak datasets were physically overlapped between the IRF and RUNX motifs, between the TEAD4 and AP1 motifs, and between USF and ATF3 (AP1) recognition motifs.

3.3.2.4 Tissue-specific preferences of predicted cofactors relative to primary TFs

Several cofactors were recurrently associated with different primary TF partners, notably in specific cell lines. One possible explanation is that these cofactors are coordinately regulated with different primary TFs preferentially in specific cell types. For example, the datasets of 25 primary TFs in which the IRF family was discovered as a cofactor were all derived from lymphoblastoid (e.g. GM12878) cell lines, with 4 exceptions (Table 3.1). Regulation by the IRF family is central to B-lymphocyte expression programs (105). All the datasets of 11 primary TFs from which the GATA and GATA1-TAL1 motifs emerged as cofactors were derived from K562 erythrocytic leukemia cells (Table 3.1),

which is consistent with the activation role that the GATA family exhibits in hematopoietic lineage gene expression (106, 107). Similarly, FOXA family members bind to the same sequences as 7 primary TFs in the HepG2 cell line derived from hepatocellular carcinoma cells (Table 3.1), which is consistent with the fact that FOXA proteins regulate the initiation of liver development (108). Datasets of GATA3 and P300 from the T47D breast cancer cell line are also linked to FOXA. Another TF family known to be a key factor regulating hepatocyte differentiation and liver-specific functions is HNF4 (109), which was discovered as a cofactor of SP1 in a HepG2 dataset. SOX2 and the SOX2-OCT4 complex were unveiled as cofactors only in datasets of 3 primary TFs from the H1-hESC cell line representing embryonic stem cells (Table 3.1), which is supported by the requirement for SOX2, OCT4 and NANOG to maintain pluripotency (110). Interestingly, all the datasets (n=12) in which YY was revealed as a cofactor were from K562 cells, with one exception (Table 3.1). Unlike the GATA TFs, the YY family is ubiquitously distributed and not known to play an especially central role in erythroid lineage development, although YY1 is known to act as a developmental repressor of the ε-globin gene along with GATA1 (111).

Not surprisingly, the SP family was found to be capable of interacting with the maximum number of TFs, which is consonant with its role in constitutive transcriptional activation. Similarly, the ubiquitously expressed AP1 interacts with 10 TFs in multiple cell lines, and these interactions do not show any preference in cell type.

A number of primary TFs exhibit an extensive capability of interacting with multiple cofactors in different tissues. The unique distribution of these cofactors across multiple cell lines suggests the tissue-specific functions of the primary TFs. For instance, TEAD4 was found to coimmunoprecipitate with GATA1-TAL1 in K562 cells, NRSF in A549 cells, FOXA in HepG2 cells, and AP1 in multiple cell types. Cofactors of P300 include IRF-RUNX in GM12878 cells, SP in H1-hESC cells, AP1 and CEBPB in HeLa-S3 cells, FOXA in HepG2 and T47D cells and GATA1-TAL1 in K562 cells. Cosegregation analysis revealed interactions between BCL11A and IRF-RUNX in GM12878 cells, and SOX2-OCT4 in H1-hESC cells. STAT5A and TBL1XR1 cosegregated with members of the IRF family in GM12878 cells and with GATA1-TAL1 in K562 cells.

3.3.2.5 Discordance between iPWMs derived from the same ChIP-seq assay

We noticed some discrepancies between IDR-thresholded datasets and SPP datasets from the same ChIP-seq assay. For example, for the primary TF BRG1, iPWMs exclusively from SPP datasets exhibit motifs of GATA1 and AP1; IDR-thresholded BRG1 data produced only noisy low information content motifs. We also noticed that the motifs derived from different biological replicates of the same ChIP-seq assay were sometimes inconsistent. One replicate of the TEAD4 ChIP-seq assay from the A549 cell line revealed only the NRSF binding motif, whereas both the cofactor AP1 and the primary motif were derived from the other replicate.

3.3.3 Novel binding motifs

We uncovered 6 high-confidence novel motifs that have not been previously annotated (Figure 3.3 – panel B). The "NM1" motif was considerably enriched in the datasets of BAF155 and BRG1 (which do not bind DNA directly) from HeLa-S3 cells and the "NM2" motif was highly conserved in the datasets of BCL11A and NANOG from H1-hESC cells. The "NM3" motif was revealed in the ESRRA and SREBF2 datasets from GM12878 cells, in the MAX dataset from HCT116, in the CREB1 and GTF3C2 datasets from K562, and in the non-DNA-binding RCOR1 dataset from IMR90 cells. The Euclidean distances between these novel motifs and primary motifs are dissimilar, ranging from 3.1 to 3.4 bits/nt. The "NM4", "NM5" and "NM6" motifs were discovered in the datasets of GATA3, MXI1 and FOSL1 from MCF-7, SK-N-SH, and H1-hESC cells, respectively, with distances ranging from 2.9 to 3.4 bits/nt.

We investigated whether these novel motifs were enriched in hallmarks of open chromatin, based on the co-occurrence with DNase I hypersensitive sites and near H3K4me and H3K27ac histone modifications (112). After scanning the complete genome with these iPWMs, the proportions of sites detected within these corresponding ENCODE chromatin tracks were determined for the respective cell lines (Table 3.2). These proportions (5%-35%) are consistent with previously reports of binding sites for other TFs (113). The frequencies of sites detected with the NM2 and NM6 motifs within

Table 3.2: Percentages of binding sites from novel motifs (NM) that overlap DNase I hypersensitive intervals and/or regions of specific histone modifications

Novel motif	ENCODE Genome Browser Track				
	DNase I HS	**H3K4me1**	**H3K4me2**	**H3K4me3**	**H3K27ac**
NM1†	4.50%	17.63%	15.52%	16.23%	11.44%
NM2†	7.06%	33.63%	14.39%	9.61%	34.05%
NM3†	4.21%	21.19%	16.89%	13.75%	12.25%
NM4	3.18%	N/A*	N/A*	1.04%	2.22%
NM5	2.31%	N/A*	N/A*	1.21%	N/A*
NM6	6.16%	32.37%	13.58%	9.36%	34.10%

† The iPWMs of the NM1, NM2 and NM3 motifs used to scan the hg19 genome assembly come from the datasets of BAF155, NANOG and ESRRA, respectively.
* The histone modification data for the specific cell line used to derive the iPWM is unavailable.

the H3K4me1 and H3K27ac peaks are significantly higher than those found after intersection of each NM binding site with the H3K4me2 and H3K4me3 tracks, respectively. The co-occurence of NM2 and NM6 with the H3K4me1 and H3K27ac epigenetic marks supports the assignment of these motifs as components of transcriptional enhancer elements, because these histone modifications are present in nucleosomes flanking enhancer elements (114). Additionally, the co-occurence of these two motifs within DNase I hypersensitive intervals exhibit the highest among all the 6 motifs. The remaining motifs could represent binding motifs of currently unknown TFs or other non-annotated functional elements.

3.3.4 Binding site motif validation

3.3.4.1 Detection of true binding sites with iPWMs

803 experimentally-confirmed, previously published binding sites were verified for the 93 TFs whose primary binding motifs had been identified (Appendix B.6). We detected these sites with the derived iPWMs by scanning promoters of known TF target genes for

binding elements with positive R_i values. There was complete concordance between these true binding sites and those detected with the iPWMs, both in terms of their locations and relative strengths. For example, an EMSA analysis of the SERPINA3 promoter proved that the nucleotide sequence starting at GRCh38 (chr14:94612260) contains a stronger binding site of STAT1 than the one starting at GRCh38 (chr14:94612291) (Appendix B.6) (115); the binding site (5'-TTCTGGTAA-3' with R_i = 9.02 bits; Row 781) detected by the bipartite iPWM is indeed 22.13 (or 4.38) fold stronger than the other site (5'-TTCTCGGA-3' with R_i = 6.89 bits; Row 782) detected in this promoter.

3.3.4.2 Correspondence between functionally characterized SNPs and changes in information content

Based on the change in the R_i value of a binding site, the effect of a SNP on the binding site strength can be predicted with iPWMs (10,12). For 153 SNPs within the binding sites of 29 TFs, we determined R_i values of the variant sequence for the corresponding iPWM and compared the predicted consequence to observed TF binding, and if available, published changes in expression (Appendix B.7). For 130 SNPs (~85.0%) affecting binding sites of 27 TFs, the predictions of the iPWMs and the experimental observations are completely concordant. For 16 SNPs (~10.5%) affecting binding sites of 10 TFs, the predicted and observed experimental findings are concordant, but the extents of these changes differ (e.g. TF binding is predicted to only be weakened, but binding or expression was completely abolished). For 7 SNPs (~4.6%) altering binding sites of 3 TFs, the predicted and observed experimental changes were discordant. iPWMs for 2 (CEBPB and SP1) of these 3 TFs were validated for other SNPs.

3.3.4.3 Comparison between iPWMs and other binding motifs

Binding motifs of eukaryotic TFs in the CIS-BP database were previously reconstructed from oligonucleotide binding selection assays (13); these motifs represent another type of ground truth reflecting the genuine sequence preferences of these TFs. For 133 TFs, we quantitatively compared the iPWMs with these motifs by determining the normalized Euclidean distances between them, and classified the distances into three categories. We

observed that the iPWMs derived in this study and the reconstructed motifs are nearly identical (<1 bit/nt) for 75 TFs, or only differ at 1 or 2 positions (1-2 bits/nt) for 18 TFs. The discovery of cofactors was the predominant explanation for large distances (>2 bits/nt) for 39 of these TFs.

3.3.4.4 Statistical analyses on iPWMs

To distinguish true binding motifs from noise motifs, the relationship between R_i values and binding energy was evaluated by performing F tests on all binding sites in all of the contiguous iPWMs that we derived (674 primary/cofactor, 312 noise). The F values are plotted as a histogram to illustrate probability density distributions (Figure 3.6; data available in Appendix B.2&B.3). The histogram shows that most F values between 0 and 100 were significantly enriched for noise motifs. In general, the F values of primary/cofactor motifs significantly exceed those derived from noise. The primary/cofactor motif and noise motif distributions are different (Mann-Whitney U test; p = 3.1E-57 at 1% significance level). We note that only primary and cofactor motifs exhibit F values >1000, which comprise 37.2% (251 of 674) of all iPWMs. The iPWMs with F values <1000 remain valid based on the other criteria described above.

3.4 Discussion

In this study, we derived and validated TF binding motifs from ChIP-seq datasets using an information theory-based approach, also revealing TF cofactor binding sites and other novel motifs. The primary TF motifs were validated by comparison with motifs derived independently from binding studies, by analysis of gene variants known to alter TF binding affinities, and by comparing the locations of binding sites predicted by iPWMs with those of true sites previously determined in published binding and expression studies. In addition to contiguous iPWMs, bipartite iPWMs with variable-length spacers were also derived. These iPWMs more precisely reflect the binding behavior of dimeric TFs, as they incorporate intermediate and often weak binding sites that are often excluded from consensus sequence-based (strong) binding site sets (3). This enables these iPWMs to accurately quantify binding site strengths across a broad range of affinities (Appendix B.6). To test this, the iPWMs were applied to mutation analyses of regulatory SNPs

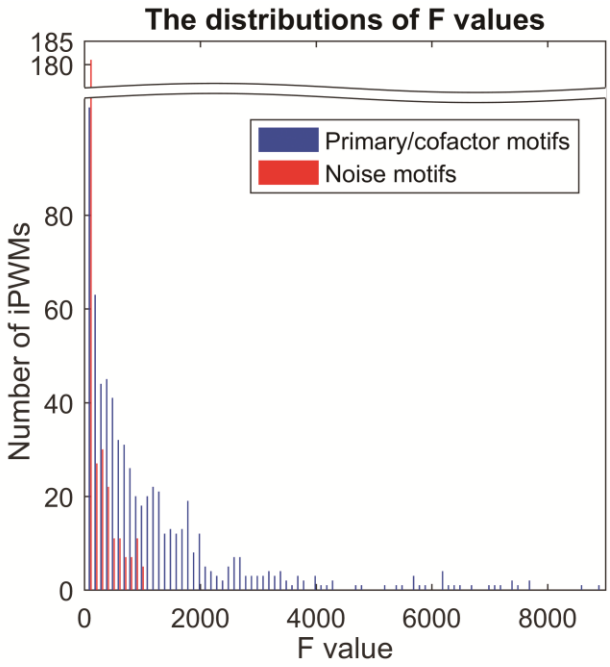

Figure 3.6: F-test results evaluating the relationship between Ri values and binding energy. The proportion of F values within the first bin for primary/cofactor motifs is much higher than that for noise motifs. A minimum threshold of 1,000 correctly classifies all the noise motifs and 37.2% (251/674) of primary/cofactor motifs.

(Appendix B.7). We have recently used this approach to identify and prioritize variants affecting TF binding in 20 risk genes of 287 hereditary breast and ovarian cancer patients (116) and 7 genes from 102 such patients (117). In present study, the iPWMs were also used to delineate known and novel TF-cofactor interactions.

TF binding sites across the genome have been predicted from promoter accessibility analyses with high-throughput DNase-seq assays. For each of 20 TFs, Yardımcı et al. (118) obtained a set of true binding sites by intersecting ChIP-seq peaks with the 50,000 strongest binding sites predicted by JASPAR and TRANSFAC PWMs in the genome. The FLR (Footprint Log-likelihood Ratio), which is defined as the logarithm of the ratio between probabilities that a DNase I footprint is produced by either a true binding site or a background sequence, was determined at these sites. We attempted to detect these true sites using the derived iPWMs. For these 20 TFs, all of these sites (ranging from n=31 to 21550, depending on the TF) were successfully detected by the iPWMs ($R_i > 0$). By

contrast, the FLR identified 35%-85% of the verified binding sites (Appendix B.8). As weak binding sites tend not to generate footprints and thus not to be discovered by DNase-seq, the expectation is that the sites detected by DNase-seq would be stronger than those that evade detection. In fact, this trend was observed for only 10 TFs and the average strengths of these classes of these binding sites were not significantly different.

In the Maskminent pipeline, the weak peaks below the threshold signal intensity do not necessarily contain weak or are missing binding sites; in fact, the distribution of R_i values of binding sites in these bottom peaks is similar to that in the top peaks used to derive the iPWM (Appendix B.1). Thresholding the dataset is required in order to ensure that the iPWM for the primary motif consists of binding sites from as many peaks as possible, while preventing alternative motifs from dominating the objective function used in Maskminent.

We also compared results produced by the Maskminent pipeline with other motif discovery tools from two perspectives of revealing primary and cofactor binding motifs (Appendix B.9). MEME-ChIP was previously used to derive motifs for 457 ChIP-seq datasets (119) and SeqGL (120) was used to analyze 105 datasets. Among the sequence-specific TFs (n=98) investigated by both tools, Maskminent and MEME-ChIP discovered primary motifs for 80 (~81.6%) and 92 (~93.9%) TFs, respectively. Among the 59 TF datasets analyzed by Maskminent, MEME-ChIP, SeqGL and HOMER (121), primary motifs were revealed for 45 (~76.3%), 51 (~86.4%), 49 (~83.1%) and 47 (~79.7%) datasets, respectively. The cofactor motifs that Maskminent found (which MEME-ChIP and SeqGL failed to detect) primarily comprise the SP family. Since MEME and SeqGL discriminate binding sites from background sequences using nucleotide frequencies computed from all input sequences, binding motifs with compositions similar to the background may fail to be discovered, such as the SP motif; in contrast, Maskminent does not rely on background compositions and will always return the lowest entropy motif. While MEME-ChIP and SeqGL revealed a greater number of cofactor motifs, selecting only the top 500 or 2000 peaks increases the likelihood that those cofactors appeared by chance. This is because MEME-ChIP and SeqGL were configured to report multiple motifs, whereas the main objective of Maskminent was to discover primary

motifs (i.e. if the initial iPWM derived from a dataset exhibits the primary motif, the masking and thresholding techniques will no longer be used, unless it is explicitly masked). Finally, the ability of Maskminent, MEME-ChIP, SeqGL to reveal binding motifs was compared on the 105 datasets (120). Each tool discovers cofactor motifs that others do not recognize.

Arvey et al. (26) trained support vector machines (SVMs) that use flexible *k*-mer patterns to capture DNA sequence signals more accurately from 286 ChIP-seq experiments than traditional motif approaches, and these SVMs can also integrate histone modifications and DNase accessibility to significantly more accurately predict TF occupancy than simpler approaches. However, the SVM approach does not provide any insight into binding strength. Even though accessible constrains increase the accuracy of binding site detection, it is not possible to compare binding site strengths once the designated sites are combined with DNase I hypersensitivity profiles and other chromatin accessibility marks.

In fact, the number of TFs for which cofactor motifs were revealed exceeds the number of TFs whose primary binding motifs were discovered, partially because only cofactor motifs can be found in the datasets of TFs which exhibit little or no sequence specificity (e.g. CCNT2, INI1 and P300). For 11 primary TFs, the binding site sequences were extremely variable; that is, the overall conservation levels of their binding motifs contain less information than noisy, low complexity sequences or cofactor motifs. For 18 primary TFs associated with cofactors, which themselves physically contact DNA, the primary TF motif was not enriched. The inability of the software to discover such primary motifs is a limitation of this approach. Interactions between the primary TFs and a subset of the cofactors which are known to cooperate with them were detected, since the association has to occur with a prevalence sufficient to produce a recognizable motif (usually >0.5 bit/nt over the entire site). Nevertheless, the algorithm may not find cofactors with weakly conserved motifs or those that overlap with other conserved motifs.

While unable to discover cofactors nor identify bipartite motifs of variable spacing, the oligonucleotide microarray technique adopted by Weirauch et al. (13) and Jolma et al. (14) theoretically is able to determine binding specificities for all the sequence-specific

TFs, because contiguous binding sites of TFs are reconstructed from overlapping oligonucleotide sequences by directly detecting complexes with the TF. This eliminates interference of noisy sequences or cofactors which may emerge as false minimum entropies using our method.

The Maskminent pipeline can be applied to other ChIP-seq data not included in ENCODE. The quality control criteria we described are capable of ensuring that the user-built iPWMs are accurate and can be used for binding site detection. The first and second criteria are particularly important, because they provide a straightforward assessment of iPWM performance. The recursively thresholded feature is crucial for guaranteeing that the discovered cofactors do not appear by chance, because the greater the number of peaks from which a cofactor is derived, the higher the confidence that the cofactor indeed interacts with the primary factor.

In summary, we comprehensively investigated and implemented a new approach to define TF binding specificities based on the ChIP-seq TF data that ENCODE has released. This allowed us to mine and quantify both known and previously unrecognized TF binding motifs and cofactor interactions on a genome scale. This information expands the granularity of the current knowledge on TF interaction with DNA and points out potential directions for future experimental study on interaction between TFs.

3.5 References

1. Leung,K.K., Ng,L.J., Ho,K.K., Tam,P.P. and Cheah,K.S. (1998) Different cis-regulatory DNA elements mediate developmental stage- and tissue-specific expression of the human COL2A1 gene in transgenic mice. J. Cell Biol., 141, 1291–1300.

2. Levine,M. and Tjian,R. (2003) Transcription regulation and animal diversity. Nature, 424, 147–151.

3. Wang,J., Zhuang,J., Iyer,S., Lin,X., Whitfield,T.W., Greven,M.C., Pierce,B.G., Dong,X., Kundaje,A., Cheng,Y., et al. (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res., 22, 1798–1812.

4. Jolma,A., Yin,Y., Nitta,K.R., Dave,K., Popov,A., Taipale,M., Enge,M., Kivioja,T., Morgunova,E. and Taipale,J. (2015) DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature, 527, 384–388.

5. Parelho,V., Hadjur,S., Spivakov,M., Leleu,M., Sauer,S., Gregson,H.C., Jarmuz,A., Canzonetta,C., Webster,Z., Nesterova,T., et al. (2008) Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell, 132, 422–433.

6. Fleming,J.D., Pavesi,G., Benatti,P., Imbriano,C., Mantovani,R. and Struhl,K. (2013) NF-Y coassociates with FOS at promoters, enhancers, repetitive elements, and inactive chromatin regions, and is stereo-positioned with growth-controlling transcription factors. Genome Res., 23, 1195–1209.

7. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature, 489, 57–74.

8. Bi,C. and Rogan,P.K. (2004) Bipartite pattern discovery by entropy minimization-based multiple local alignment. Nucleic Acids Res., 32, 4979–4991.

9. Shannon,C.E. (1948) A Mathematical Theory of Communication. Bell Syst. Technol. J., 27, 379–423, 623–656.

10. Schneider,T.D. (1997) Information content of individual genetic sequences. J. Theor. Biol., 189, 427–441.

11. Schneider,T.D. (1999) Measuring molecular information. J. Theor. Biol., 201, 87–92.

12. Rogan,P.K., Faux,B.M. and Schneider,T.D. (1998) Information analysis of human splice site mutations. Hum. Mutat., 12, 153-171 .

13. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K., et al. (2014) Determination and inference of eukaryotic transcription factor sequence specificity. Cell, 158, 1431–1443.

14. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G., et al. (2013) DNA-binding specificities of human transcription factors. Cell, 152, 327–339.

15. Kheradpour,P. and Kellis,M. (2014) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res., 42, 2976–2987.

16. Schneider,T.D. (2002) Consensus sequence Zen. Appl. Bioinformatics, 1, 111–119.

17. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P., et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res., 22, 1813–1831.

18. Wang,J., Zhuang,J., Iyer,S., Lin,X.-Y., Greven,M.C., Kim,B.-H., Moore,J., Pierce,B.G., Dong,X., Virgil,D., et al. (2013) Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. Nucleic Acids Res., 41, D171-176.

19. Kharchenko,P.V., Tolstorukov,M.Y. and Park,P.J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat. Biotechnol., 26, 1351–1359.

20. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res., 37, W202-208.

21. Tanay,A. (2006) Extensive low-affinity transcriptional interactions in the yeast genome. Genome Res., 16, 962–972.

22. Marzouk,S. El, Gahattamaneni,R., Joshi,S.R. and Scovell,W.M. (2008) The plasticity of estrogen receptor-DNA complexes: binding affinity and specificity of estrogen receptors to estrogen response element half-sites separated by variant spacers. J. Steroid Biochem. Mol. Biol., 110, 186–195.

23. Eckert,D., Buhl,S., Weber,S., Jager,R. and Schorle,H. The AP-2 family of transcription factors. Genome Biol., 6, 2005.

24. Ehret,G.B., Reichenbach,P., Schindler,U., Horvath,C.M., Fritz,S., Nabholz,M. and Bucher,P. (2001) DNA binding specificity of different STAT proteins. Comparison of *in vitro* specificity with natural target sites. J. Biol. Chem., 276, 6675–6688.

25. Kataoka,K., Noda,M. and Nishizawa,M. (1994) Maf nuclear oncoprotein recognizes sequences related to an AP-1 site and forms heterodimers with both Fos and Jun. Mol. Cell. Biol., 14, 700–712.

26. Arvey,A., Agius,P., Noble,W.S. and Leslie,C. (2012) Sequence and chromatin determinants of cell-type-specific transcription factor binding. Genome Res., 22, 1723–1734.

27. Kawana,M., Lee,M.E., Quertermous,E.E. and Quertermous,T. (1995) Cooperative interaction of GATA-2 and AP1 regulates transcription of the endothelin-1 gene. Mol. Cell. Biol., 15, 4225–4231.

28. Roca,H., Pande,M., Huo,J.S., Hernandez,J., Cavalcoli,J.D., Pienta,K.J. and McEachin,R.C. (2014) A bioinformatics approach reveals novel interactions of the OVOL transcription factors in the regulation of epithelial - mesenchymal cell reprogramming and cancer progression. BMC Syst. Biol., 8, 29.

29. Zhu,C., Johansen,F.E. and Prywes,R. (1997) Interaction of ATF6 and serum response factor. Mol. Cell. Biol., 17, 4957–4966.

30. Zhang,X., Wrzeszczynska,M.H., Horvath,C.M. and Darnell,J.E. (1999) Interacting regions in Stat3 and c-Jun that participate in cooperative transcriptional activation. Mol. Cell. Biol., 19, 7138–7146.

31. Ito,T., Yamauchi,M., Nishina,M., Yamamichi,N., Mizutani,T., Ui,M., Murakami,M. and Iba,H. (2001) Identification of SWI.SNF complex subunit BAF60a as a determinant of the transactivation potential of Fos/Jun dimers. J. Biol. Chem., 276, 2852–2857.

32. Na,S.Y., Choi,J.E., Kim,H.J., Jhun,B.H., Lee,Y.C. and Lee,J.W. (1999) Bcl3, an IkappaB protein, stimulates activating protein-1 transactivation and cellular proliferation. J. Biol. Chem., 274, 28491–28496.

33. Henderson,A., Holloway,A., Reeves,R. and Tremethick,D.J. (2004) Recruitment of SWI/SNF to the human immunodeficiency virus type 1 promoter. Mol. Cell. Biol., 24, 389–397.

34. Lee,J.S., See,R.H., Deng,T. and Shi,Y. (1996) Adenovirus E1A downregulates cJun- and JunB-mediated transcription by targeting their coactivator p300. Mol. Cell. Biol., 16, 4312–4326.

35. Schwartz,C., Beck,K., Mink,S., Schmolke,M., Budde,B., Wenning,D. and Klempnauer,K.-H. (2003) Recruitment of p300 by C/EBPbeta triggers phosphorylation of p300 and modulates coactivator activity. EMBO J., 22, 882–892.

36. Bailey,S.D., Zhang,X., Desai,K., Aid,M., Corradin,O., Cowper-Sal Lari,R., Akhtar-Zaidi,B., Scacheri,P.C., Haibe-Kains,B. and Lupien,M. (2015) ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. Nat. Commun., 2, 6186.

37. O'Geen,H., Lin,Y.-H., Xu,X., Echipare,L., Komashko,V.M., He,D., Frietze,S., Tanabe,O., Shi,L., Sartor,M.A., et al. (2010) Genome-wide binding of the orphan nuclear receptor TR4 suggests its general role in fundamental biological processes. BMC Genomics, 11, 689.

38. Elagib,K.E., Racke,F.K., Mogass,M., Khetawat,R., Delehanty,L.L. and Goldfarb,A.N. (2003) RUNX1 and GATA-1 coexpression and cooperation in megakaryocytic differentiation. Blood, 101, 4333–4341.

39. Xu,Z., Meng,X., Cai,Y., Koury,M.J. and Brandt,S.J. (2006) Recruitment of the SWI/SNF protein Brg1 by a multiprotein complex effects transcriptional repression in murine erythroid progenitors. Biochem. J., 399, 297–304.

40. Grau,J., Grosse,I., Posch,S. and Keilwagen,J. (2015) Motif clustering with implications for transcription factor interactions. Ger. Conf. Bioinforma.

41. Albergaria,A., Paredes,J., Sousa,B., Milanezi,F., Carneiro,V., Bastos,J., Costa,S., Vieira,D., Lopes,N., Lam,E.W., et al. (2009) Expression of FOXA1 and GATA-3 in breast cancer: the prognostic significance in hormone receptor-negative tumours. Breast Cancer Res. 11, R40.

42. Cirillo,L.A. and Zaret,K.S. (1999) An early developmental transcription factor complex that is more stable on nucleosome core particles than on free DNA. Mol. Cell, 4, 961–969.

43. Grabowska,M.M., Elliott,A.D., DeGraff,D.J., Anderson,P.D., Anumanthan,G., Yamashita,H., Sun,Q., Friedman,D.B., Hachey,D.L., Yu,X., et al. (2014) NFI transcription factors interact with FOXA1 to regulate prostate-specific gene expression. Mol. Endocrinol. Baltim. Md, 28, 949–964.

44. Kohler,S. and Cirillo,L.A. (2010) Stable chromatin binding prevents FoxA acetylation, preserving FoxA chromatin remodeling. J. Biol. Chem., 285, 464–472.

45. Kardassis,D., Falvey,E., Tsantili,P., Hadzopoulou-Cladaras,M. and Zannis,V. (2002) Direct physical interactions between HNF-4 and Sp1 mediate synergistic transactivation of the apolipoprotein CIII promoter. Biochemistry (Mosc.), 41, 1217–1228.

46. Xu,L., Ma,X., Bagattin,A. and Mueller,E. (2016) The transcriptional coactivator PGC1α protects against hyperthermic stress via cooperation with the heat shock factor HSF1. Cell Death Dis., 7, e2102.

47. Hurgin,V., Novick,D. and Rubinstein,M. (2002) The promoter of IL-18 binding protein: activation by an IFN-gamma -induced complex of IFN regulatory factor 1 and CCAAT/enhancer binding protein beta. Proc. Natl. Acad. Sci. U. S. A., 99, 16957–16962.

48. Kuwata,T., Gongora,C., Kanno,Y., Sakaguchi,K., Tamura,T., Kanno,T., Basrur,V., Martinez,R., Appella,E., Golub,T., et al. (2002) Gamma interferon triggers interaction

between ICSBP (IRF-8) and TEL, recruiting the histone deacetylase HDAC3 to the interferon-responsive element. Mol. Cell. Biol., 22, 7439–7448.

49. Leng,R.-X., Wang,W., Cen,H., Zhou,M., Feng,C.-C., Zhu,Y., Yang,X.-K., Yang,M., Zhai,Y., Li,B.-Z., et al. (2012) Gene-gene and gene-sex epistatic interactions of MiR146a, IRF5, IKZF1, ETS1 and IL21 in systemic lupus erythematosus. PloS One, 7, e51090.

50. Drew,P.D., Franzoso,G., Becker,K.G., Bours,V., Carlson,L.M., Siebenlist,U. and Ozato,K. (1995) NF kappa B and interferon regulatory factor 1 physically interact and synergistically induce major histocompatibility class I gene expression. J. Interferon Cytokine Res. Off. J. Int. Soc. Interferon Cytokine Res., 15, 1037–1045.

51. Ziegler-Heitbrock,L., Lötzerich,M., Schaefer,A., Werner,T., Frankenberger,M. and Benkhart,E. (2003) IFN-alpha induces the human IL-10 gene by recruiting both IFN regulatory factor 1 and Stat3. J. Immunol. Baltim. Md 1950, 171, 285–290.

52. Dornan,D., Eckert,M., Wallace,M., Shimizu,H., Ramsay,E., Hupp,T.R. and Ball,K.L. (2004) Interferon regulatory factor 1 binding to p300 stimulates DNA-dependent acetylation of p53. Mol. Cell. Biol., 24, 10083–10098.

53. Roopra,A., Sharling,L., Wood,I.C., Briggs,T., Bachfischer,U., Paquette,A.J. and Buckley,N.J. (2000) Transcriptional repression by neuron-restrictive silencer factor is mediated via the Sin3-histone deacetylase complex. Mol. Cell. Biol., 20, 2147–2157.

54. Gutierrez,S., Javed,A., Tennant,D.K., van Rees,M., Montecino,M., Stein,G.S., Stein,J.L. and Lian,J.B. (2002) CCAAT/enhancer-binding proteins (C/EBP) beta and delta activate osteocalcin gene transcription and synergize with Runx2 at the C/EBP element to regulate bone-specific expression. J. Biol. Chem., 277, 1316–1323.

55. Ciavatta,D.J., Yang,J., Preston,G.A., Badhwar,A.K., Xiao,H., Hewins,P., Nester,C.M., Pendergraft,W.F., Magnuson,T.R., Jennette,J.C., et al. (2010) Epigenetic basis for aberrant upregulation of autoantigen genes in humans with ANCA vasculitis. J. Clin. Invest., 120, 3209–3219.

56. Kitabayashi,I., Yokoyama,A., Shimizu,K. and Ohki,M. (1998) Interaction and functional cooperation of the leukemia-associated factors AML1 and p300 in myeloid cell differentiation. EMBO J., 17, 2994–3004.

57. Chiang,B.-T., Liu,Y.-W., Chen,B.-K., Wang,J.-M. and Chang,W.-C. (2006) Direct interaction of C/EBPdelta and Sp1 at the GC-enriched promoter region synergizes the IL-10 gene transcription in mouse macrophage. J. Biomed. Sci., 13, 621–635.

58. Höcker,M., Raychowdhury,R., Plath,T., Wu,H., O'Connor,D.T., Wiedenmann,B., Rosewicz,S. and Wang,T.C. (1998) Sp1 and CREB mediate gastrin-dependent regulation of chromogranin A promoter activity in gastric carcinoma cells. J. Biol. Chem., 273, 34000–34007.

59. Syddall,C.M., Reynard,L.N., Young,D.A. and Loughlin,J. (2013) The identification of trans-acting factors that regulate the expression of GDF5 via the osteoarthritis susceptibility SNP rs143383. PLoS Genet., 9, e1003557.

60. Karlseder,J., Rotheneder,H. and Wintersberger,E. (1996) Interaction of Sp1 with the growth- and cell cycle-regulated transcription factor E2F. Mol. Cell. Biol., 16, 1659–1667.

61. Corominas,R., Yang,X., Lin,G.N., Kang,S., Shen,Y., Ghamsari,L., Broly,M., Rodriguez,M., Tam,S., Trigg,S.A., et al. (2014) Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. Nat. Commun., 5, 3650.

62. Hu,X., Li,T., Zhang,C., Liu,Y., Xu,M., Wang,W., Jia,Z., Ma,K., Zhang,Y. and Zhou,C. (2011) GATA4 regulates ANF expression synergistically with Sp1 in a cardiac hypertrophy model. J. Cell. Mol. Med., 15, 1865–1877.

63. Xie,R.-L., Gupta,S., Miele,A., Shiffman,D., Stein,J.L., Stein,G.S. and van Wijnen,A.J. (2003) The tumor suppressor interferon regulatory factor 1 interferes with SP1 activation to repress the human CDK2 promoter. J. Biol. Chem., 278, 26589–26596.

64. Kyo,S., Takakura,M., Taira,T., Kanaya,T., Itoh,H., Yutsudo,M., Ariga,H. and Inoue,M. (2000) Sp1 cooperates with c-Myc to activate transcription of the human telomerase reverse transcriptase gene (hTERT). Nucleic Acids Res., 28, 669–677.

65. Gartel,A.L., Ye,X., Goufman,E., Shianov,P., Hay,N., Najmabadi,F. and Tyner,A.L. (2001) Myc represses the p21(WAF1/CIP1) promoter and interacts with Sp1/Sp3. Proc. Natl. Acad. Sci. U. S. A., 98, 4510–4515.

66. Natesampillai,S., Fernandez-Zapico,M.E., Urrutia,R. and Veldhuis,J.D. (2006) A novel functional interaction between the Sp1-like protein KLF13 and SREBP-Sp1 activation complex underlies regulation of low density lipoprotein receptor promoter function. J. Biol. Chem., 281, 3040–3047.

67. Carver,B.J., Plosa,E.J., Stinnett,A.M., Blackwell,T.S. and Prince,L.S. (2013) Interactions between NF-κB and SP3 connect inflammatory signaling with reduced FGF-10 expression. J. Biol. Chem., 288, 15318–15325.

68. Roder,K., Wolf,S.S., Larkin,K.J. and Schweizer,M. (1999) Interaction between the two ubiquitously expressed transcription factors NF-Y and Sp1. Gene, 234, 61–69.

69. Smith,K.T., Coffee,B. and Reines,D. (2004) Occupancy and synergistic activation of the FMR1 promoter by Nrf-1 and Sp1 *in vivo*. Hum. Mol. Genet., 13, 1611–1621.

70. Formisano,L., Guida,N., Valsecchi,V., Cantile,M., Cuomo,O., Vinciguerra,A., Laudati,G., Pignataro,G., Sirabella,R., Di Renzo,G., et al. (2015) Sp3/REST/HDAC1/HDAC2 Complex Represses and Sp1/HIF-1/p300 Complex Activates ncx1 Gene Transcription, in Brain Ischemia and in Ischemic Brain Preconditioning, by Epigenetic Mechanism. J. Neurosci. Off. J. Soc. Neurosci., 35, 7332–7348.

71. Ingram,R.M., Valeaux,S., Wilson,N., Bouhlel,M.A., Clarke,D., Krüger,I., Kulu,D., Suske,G., Philipsen,S., Tagoh,H., et al. (2011) Differential regulation of sense and antisense promoter activity at the Csf1R locus in B cells by the transcription factor PAX5. Exp. Hematol., 39, 730-740–2.

72. Giannopoulou,E.G. and Elemento,O. (2013) Inferring chromatin-bound protein complexes from genome-wide binding assays. Genome Res., 23, 1295–1306.

73. Chakravarty,K., Wu,S.-Y., Chiang,C.-M., Samols,D. and Hanson,R.W. (2004) SREBP-1c and Sp1 interact to regulate transcription of the gene for phosphoenolpyruvate carboxykinase (GTP) in the liver. J. Biol. Chem., 279, 15385–15395.

74. Lim,K. and Chang,H.-I. (2010) O-GlcNAc inhibits interaction between Sp1 and sterol regulatory element binding protein 2. Biochem. Biophys. Res. Commun., 393, 314–318.

75. Biesiada,E., Hamamori,Y., Kedes,L. and Sartorelli,V. (1999) Myogenic basic helix-loop-helix proteins and Sp1 interact as components of a multiprotein transcriptional complex required for activity of the human cardiac alpha-actin promoter. Mol. Cell. Biol., 19, 2577–2584.

76. Look,D.C., Pelletier,M.R., Tidwell,R.M., Roswit,W.T. and Holtzman,M.J. (1995) Stat1 depends on transcriptional synergy with Sp1. J. Biol. Chem., 270, 30264–30267.

77. Rossi,A., Mukerjee,R., Ferrante,P., Khalili,K., Amini,S. and Sawaya,B.E. (2006) Human immunodeficiency virus type 1 Tat prevents dephosphorylation of Sp1 by TCF-4 in astrocytes. J. Gen. Virol., 87, 1613–1623.

78. Kim,E., Yang,Z., Liu,N.-C. and Chang,C. (2005) Induction of apolipoprotein E expression by TR4 orphan nuclear receptor via 5’ proximal promoter region. Biochem. Biophys. Res. Commun., 328, 85–90.

79. Lee,J.S., Galvin,K.M. and Shi,Y. (1993) Evidence for physical interaction between the zinc-finger transcription factors YY1 and Sp1. Proc. Natl. Acad. Sci. U. S. A., 90, 6145–6149.

80. Lee,D.-K., Suh,D., Edenberg,H.J. and Hur,M.-W. (2002) POZ domain transcription factor, FBI-1, represses transcription of ADH5/FDH by interacting with the zinc finger and interfering with DNA binding activity of Sp1. J. Biol. Chem., 277, 26761–26768.

81. Abramovitch,S., Glaser,T., Ouchi,T. and Werner,H. (2003) BRCA1-Sp1 interactions in transcriptional regulation of the IGF-IR gene. FEBS Lett., 541, 149–154.

82. Zhang,Y. and Dufau,M.L. (2003) Repression of the luteinizing hormone receptor gene promoter by cross talk among EAR3/COUP-TFI, Sp1/Sp3, and TFIIB. Mol. Cell. Biol., 23, 6958–6972.

83. Vallian,S., Chin,K.V. and Chang,K.S. (1998) The promyelocytic leukemia protein interacts with Sp1 and inhibits its transactivation of the epidermal growth factor receptor promoter. Mol. Cell. Biol., 18, 7147–7156.

84. Plaisance,V., Niederhauser,G., Azzouz,F., Lenain,V., Haefliger,J.-A., Waeber,G. and Abderrahmani,A. (2005) The repressor element silencing transcription factor (REST)-mediated transcriptional repression requires the inhibition of Sp1. J. Biol. Chem., 280, 401–407.

85. Su,D., Peng,X., Zhu,S., Huang,Y., Dong,Z., Zhang,Y., Zhang,J., Liang,Q., Lu,J. and Huang,B. (2011) Role of p38 MAPK pathway in BMP4-mediated Smad-dependent premature senescence in lung cancer cells. Biochem. J., 433, 333–343.

86. Kadam,S., McAlpine,G.S., Phelan,M.L., Kingston,R.E., Jones,K.A. and Emerson,B.M. (2000) Functional selectivity of recombinant mammalian SWI/SNF subunits. Genes Dev., 14, 2441–2451.

87. Liu,W.-L., Coleman,R.A., Ma,E., Grob,P., Yang,J.L., Zhang,Y., Dailey,G., Nogales,E. and Tjian,R. (2009) Structures of three distinct activator-TFIID complexes. Genes Dev., 23, 1510–1521.

88. Gagliardi,A., Mullin,N.P., Ying Tan,Z., Colby,D., Kousa,A.I., Halbritter,F., Weiss,J.T., Felker,A., Bezstarosti,K., Favaro,R., et al. (2013) A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. EMBO J., 32, 2231–2247.

89. Ambrosetti,D.C., Basilico,C. and Dailey,L. (1997) Synergistic activation of the fibroblast growth factor 4 enhancer by Sox2 and Oct-3 depends on protein-protein

interactions facilitated by a specific spatial arrangement of factor binding sites. Mol. Cell. Biol., 17, 6321–6329.

90. Ishiguro,A., Kassavetis,G.A. and Geiduschek,E.P. (2002) Essential roles of Bdp1, a subunit of RNA polymerase III initiation factor TFIIIB, in transcription and tRNA processing. Mol. Cell. Biol., 22, 3264–3275.

91. Bockmühl,Y., Patchev,A.V., Madejska,A., Hoffmann,A., Sousa,J.C., Sousa,N., Holsboer,F., Almeida,O.F.X. and Spengler,D. (2015) Methylation at the CpG island shore region upregulates Nr3c1 promoter activity after early-life stress. Epigenetics, 10, 247–257.

92. Chiang,C.M. and Roeder,R.G. (1995) Cloning of an intrinsic human TFIID subunit that interacts with multiple transcriptional activators. Science, 267, 531–536.

93. Zhou,Z., Li,X., Deng,C., Ney,P.A., Huang,S. and Bungert,J. (2010) USF and NF-E2 cooperate to regulate the recruitment and activity of RNA polymerase II in the beta-globin gene locus. J. Biol. Chem., 285, 15894–15905.

94. Kiryu-Seo,S., Kato,R., Ogawa,T., Nakagomi,S., Nagata,K. and Kiyama,H. (2008) Neuronal injury-inducible gene is synergistically regulated by ATF3, c-Jun, and STAT3 through the interaction with Sp1 in damaged neurons. J. Biol. Chem., 283, 6988–6996.

95. Noti,J.D. (1997) Sp3 mediates transcriptional activation of the leukocyte integrin genes CD11C and CD11B and cooperates with c-Jun to activate CD11C. J. Biol. Chem., 272, 24038–24045.

96. Lim,K. and Chang,H.-I. (2009) O-GlcNAc inhibits interaction between Sp1 and Elf-1 transcription factors. Biochem. Biophys. Res. Commun., 380, 569–574.

97. Tsai,E.Y., Falvo,J.V., Tsytsykova,A.V., Barczak,A.K., Reimold,A.M., Glimcher,L.H., Fenton,M.J., Gordon,D.C., Dunn,I.F. and Goldfeld,A.E. (2000) A lipopolysaccharide-specific enhancer complex involving Ets, Elk-1, Sp1, and CREB binding protein and p300 is recruited to the tumor necrosis factor alpha promoter *in vivo*. Mol. Cell. Biol., 20, 6084–6094.

98. Galvagni,F., Orlandini,M. and Oliviero,S. (2013) Role of the AP-1 transcription factor FOSL1 in endothelial cells adhesion and migration. Cell Adhes. Migr., 7, 408–411.

99. Rosmarin,A.G., Luo,M., Caprio,D.G., Shang,J. and Simkevich,C.P. (1998) Sp1 cooperates with the ets transcription factor, GABP, to activate the CD18 (beta2 leukocyte integrin) promoter. J. Biol. Chem., 273, 13097–13103.

100. Latinkić,B.V., Zeremski,M. and Lau,L.F. (1996) Elk-1 can recruit SRF to form a ternary complex upon the serum response element. Nucleic Acids Res., 24, 1345–1351.

101. Liu,X., Li,H., Rajurkar,M., Li,Q., Cotton,J.L., Ou,J., Zhu,L.J., Goel,H.L., Mercurio,A.M., Park,J.-S., et al. (2016) Tead and AP1 Coordinate Transcription and Motility. Cell Rep., 14, 1169–1180.

102. Stewart,M.D., Choi,Y., Johnson,G.A., Yu-Lee,L., Bazer,F.W. and Spencer,T.E. (2002) Roles of Stat1, Stat2, and interferon regulatory factor-9 (IRF-9) in interferon tau regulation of IRF-1. Biol. Reprod., 66, 393–400.

103. Wadman,I.A., Osada,H., Grütz,G.G., Agulnick,A.D., Westphal,H., Forster,A. and Rabbitts,T.H. (1997) The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. EMBO J., 16, 3145–3157.

104. Huang,S., Qiu,Y., Stein,R.W. and Brandt,S.J. (1999) p300 functions as a transcriptional coactivator for the TAL1/SCL oncoprotein. Oncogene, 18, 4958–4967.

105. Honda,K. and Taniguchi,T. (2006) IRFs: master regulators of signalling by Toll-like receptors and cytosolic pattern-recognition receptors. Nat. Rev. Immunol., 6, 644–658.

106. Ferreira,R., Ohneda,K., Yamamoto,M. and Philipsen,S. (2005) GATA1 function, a paradigm for transcription factors in hematopoiesis. Mol. Cell. Biol., 25, 1215–1227.

107. Woon Kim,Y., Kim,S., Geun Kim,C. and Kim,A. (2011) The distinctive roles of erythroid specific activator GATA-1 and NF-E2 in transcription of the human fetal γ-globin genes. Nucleic Acids Res., 39, 6944–6955.

108. Lee,C.S., Friedman,J.R., Fulmer,J.T. and Kaestner,K.H. (2005) The initiation of liver development is dependent on Foxa transcription factors. Nature, 435, 944–947.

109. Bonzo,J.A., Ferry,C.H., Matsubara,T., Kim,J.-H. and Gonzalez,F.J. (2012) Suppression of hepatocyte proliferation by hepatocyte nuclear factor 4α in adult mice. J. Biol. Chem., 287, 7345–7356.

110. Rodda,D.J., Chew,J.-L., Lim,L.-H., Loh,Y.-H., Wang,B., Ng,H.-H. and Robson,P. (2005) Transcriptional regulation of nanog by OCT4 and SOX2. J. Biol. Chem., 280, 24731–24737.

111. Raich,N., Clegg,C.H., Grofti,J., Roméo,P.H. and Stamatoyannopoulos,G. (1995) GATA1 and YY1 are developmental repressors of the human epsilon-globin gene. EMBO J., 14, 801–809.

112. Yan,C. and Boyd,D.D. (2006) Histone H3 Acetylation and H3 K4 Methylation Define Distinct Chromatin Regions Permissive for Transgene Expression. Mol. Cell. Biol., 26, 6357–6371.

113. Rye,M., Sætrom,P., Håndstad,T. and Drabløs,F. (2011) Clustered ChIP-Seq-defined transcription factor binding sites and histone modifications map distinct classes of regulatory elements. BMC Biol., 9, 80.

114. Calo,E. and Wysocka,J. (2013) Modification of enhancer chromatin: what, how, and why? Mol. Cell, 49, 825–837.

115. Kordula,T., Rydel,R.E., Brigham,E.F., Horn,F., Heinrich,P.C. and Travis,J. (1998) Oncostatin M and the interleukin-6 and soluble interleukin-6 receptor complex regulate alpha1-antichymotrypsin expression in human cortical astrocytes. J. Biol. Chem., 273, 4112–4118.

116. Caminsky,N.G., Mucaki,E.J., Perri,A.M., Lu,R., Knoll,J.H.M. and Rogan,P.K. (2016) Prioritizing Variants in Complete Hereditary Breast and Ovarian Cancer Genes in Patients Lacking Known BRCA Mutations. Hum. Mutat., 37, 640–652.

117. Mucaki,E.J., Caminsky,N.G., Perri,A.M., Lu,R., Laederach,A., Halvorsen,M., Knoll,J.H.M. and Rogan,P.K. (2016) A unified analytic framework for prioritization of non-coding variants of uncertain significance in heritable breast and ovarian cancer. BMC Med. Genomics, 9, 19.

118. Yardımcı,G.G., Frank,C.L., Crawford,G.E. and Ohler,U. (2014) Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. Nucleic Acids Res., 42, 11865–11878.

119. Machanick,P. and Bailey,T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics, 27, 1696-1697.

120. Setty,M. and Leslie,C.S. (2015) SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps. PLoS Comput. Biol., 11, e1004271.

121. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell, 38, 576–589.

Chapter 4

4 A Unified Analytic Framework for Prioritization of Non-coding Variants of Uncertain Significance in Heritable Breast and Ovarian Cancer

The work presented in this chapter is reproduced from:

Mucaki,E.J., Caminsky,N.G., Perri,A.M., Lu,R., Laederach,A., Halvorsen,M., Knoll,J.H.M., Rogan,P.K. (2016) A unified analytic framework for prioritization of non-coding variants of uncertain significance in heritable breast and ovarian cancer. *BMC Med. Genomics*, 9, 19

4.1 Background

Advances in next-generation sequencing (NGS) have enabled panels of genes, whole exomes, and even whole genomes to be sequenced for multiple individuals in parallel. These platforms have become so cost-effective and accurate that they are beginning to be adopted in clinical settings, as evidenced by recent FDA approvals [1, 2]. However, the overwhelming number of gene variants revealed in each individual has challenged interpretation of clinically significant genetic variation [3–5].

After common variants, which are rarely pathogenic, are eliminated, the number of variants of uncertain significance (VUS) in the residual set remains substantial. Assessment of pathogenicity is not trivial, considering that nearly half of the unique variants are novel, and cannot be resolved using published literature and variant databases [6]. Furthermore, loss-of-function variants (those resulting in protein truncation are most likely to be deleterious) represent a very small proportion of identified variants. The remaining variants are missense and synonymous variants in the exon, single nucleotide changes, or in frame insertions or deletions in intervening and intergenic regions. Functional analysis of large numbers of these variants often cannot be performed, due to lack of relevant tissues, and the cost, time, and labor required for each variant. Another problem is that *in silico* protein coding prediction tools exhibit inconsistent accuracy and are thus problematic for clinical risk evaluation [7–9].

Consequently, many HBOC patients undergoing genetic susceptibility testing will receive either an inconclusive (no *BRCA* variant identified) or an uncertain (*BRCA* VUS) result. The former has been reported in up to 80% of cases and depends on the number of genes tested [10]. The occurrence of uncertain *BRCA* mutations varies greatly (as high as 46% in African American populations and as low as 2.1%) among tested individuals depending on the laboratory and the patient's ethnicity [11–13]. The inconsistency in diagnostic yield is significant, considering that HBOC accounts for 5-10% of all breast/ovarian cancer [14, 15].

One strategy to improve variant interpretation in patients is to reduce the full set of variants to a manageable list of potentially pathogenic variants. Evidence for pathogenicity of VUS in genetic disease is often limited to amino acid coding changes [16, 17], and mutations affecting splicing, transcriptional activation, and mRNA stability tend to be underreported [18–24]. Splicing errors are estimated to represent 15% of disease-causing mutations [25], but may be much higher [26, 27]. The impact of a single nucleotide change in a recognition sequence can range from insignificant to complete abolition of a protein binding site. Aberrant splicing events causing frameshifts often disrupt of protein function; in-frame changes are dependent on gene context. The complexity of interpretation of non-coding sequence variants benefits from computational approaches [28] and direct functional analyses [29–33] that may each support evidence of pathogenicity.

Ex vivo transfection assays developed to determine the pathogenicity of VUS predicted to lead to splicing aberrations (using *in silico* tools) have been successful in identifying pathogenic sequence variants [34, 35]. Information technology (IT)-based analysis of splicing variants has proven to be robust and accurate (as determined by functional assays for mRNA expression or binding assays) at analyzing splice site (SS) variants, including splicing regulatory factor binding sites (SRFBSs), and in distinguishing them from polymorphisms in both rare and common diseases [36–39]. However, IT can be applied to any sequence recognized and bound by another factor [40], such as with transcription factor binding sites (TFBSs) and RNA-binding protein binding sites (RBBSs). IT is used as a measure of sequence conservation and is more accurate than consensus sequences

[41]. The individual information (R_i) of a base is related to thermodynamic entropy, and therefore free energy of binding, and is measured on a logarithmic scale (in bits). By comparing the change in information (ΔR_i) for a nucleotide variation of a bound sequence, the resulting change in binding affinity is $\geq 2\Delta R_i$, such that a 1 bit change in information will result in at least a 2-fold change in binding affinity [42].

IT measures nucleotide sequence conservation and does not provide information on effects of variants on mRNA secondary (2°) structure, nor can it accurately predict effects of amino acid sequence changes. Associations of structural changes in untranslated regions of mRNA (UTR) with disease justifies including predicted effects of these changes on 2° structure in the comprehensive analysis of sequence variants [43]. Other *in silico* methods have attempted to address these deficiencies. For example, Halvorsen et al. (2010) introduced an algorithm called SNPfold, which computes the potential effect of a single nucleotide variant (SNV) on mRNA 2° structure [20]. Predictions made by SNPfold can be tested by the SHAPE assay (Selective 2'-Hydroxyl Acylation analyzed by Primer Extension) [44], which provides evidence for sequence variants that lead to structural changes in mRNA by detection of covalent adducts in mRNA.

The implications of improved VUS interpretation are particularly relevant for HBOC due to its incidence and the adoption of panel testing for these individuals [45, 46]. It has been suggested that patients with a high risk profile receiving uninformative results would imply that deleterious variants lie in untested regions of *BRCA1/2*, untested genes, or are unrecognized [47, 48]. This is also supported by studies where families with linkage to *BRCA1/2* had no detectable pathogenic mutation (however it is noteworthy that detection rates of *BRCA* mutations in families with documented linkage to these loci appears to vary by ascertainment, inclusion criteria, and technology used to identify the mutations) [49, 50]. The concept of non-*BRCA* gene association has been demonstrated by the identification of low-to-moderate risk HBOC genes, and variants within coding and non-coding regions affecting splicing and regulatory factor binding [51, 52]. Consequently, VUS, which include rare missense changes, other coding and non-coding changes in all of these genes, greatly outnumber the catalog of known deleterious mutations [53].

Here, we develop and evaluate IT-based models to predict potential non-coding sequence mutations in SSs, TFBSs, and RBBSs in 7 genes sequenced in their entirety. These models were used to analyze 102 anonymous HBOC patients who did not exhibit known *BRCA1/2* coding mutations at the time of initial testing, despite meeting the criteria for *BRCA* genetic testing. The genes are: *ATM*, *BRCA1*, *BRCA2*, *CDH1*, *CHEK2*, *PALB2*, and *TP53*, and have been reported to harbor mutations that increase HBOC risk [54–76]. We apply these IT-based methods to analyze variants in the complete sequences of coding, non-coding, and up- and downstream regions of the 7 genes. In this study, we established and applied a unified IT-based framework, first filtering out common variants, then to "flag" potentially deleterious ones. Then, using context-specific criteria and information from the published literature, we prioritized likely candidates.

4.2 Methods

4.2.1 Design of Tiled Capture Array for HBOC Gene Panel

Nucleic acid hybridization capture reagents designed from genomic sequences generally avoid repetitive sequence content to avoid cross hybridization [77]. Complete gene sequences harbor numerous repetitive sequences, and an excess of denatured C0t-1 DNA is usually added to hybridization to prevent inclusion of these sequences [78]. RepeatMasker software completely masks all repetitive and low-complexity sequences [79]. We increased sequence coverage in complete genes with capture probes by enriching for both single-copy and divergent repeat (> 30% divergence) regions, such that, under the correct hybridization and wash conditions, all probes hybridize only to their correct genomic locations [77]. This step was incorporated into a modified version of Gnirke and colleagues' (2009) in-solution hybridization enrichment protocol, in which the majority of library preparation, pull-down, and wash steps were automated using a BioMek® FXP Automation Workstation (Beckman Coulter, Mississauga, Canada) [80].

Genes *ATM* (RefSeq: NM_000051.3, NP_000042.3), *BRCA1* (RefSeq: NM_007294.3, NP_009225.1), *BRCA2* (RefSeq: NM_000059.3, NP_000050.2), *CDH1* (RefSeq: NM_004360.3, NP_004351.1), *CHEK2* (RefSeq: NM_145862.2, NP_665861.1), *PALB2* (RefSeq: NM_024675.3, NP_078951.2), and *TP53* (RefSeq: NM_000546.5,

NP_000537.3) were selected for capture probe design by targeting single copy or highly divergent repeat regions (spanning 10 kb up- and downstream of each gene relative to the most upstream first exon and most downstream final exon in RefSeq) using an *ab initio* approach [77]. If a region was excluded by *ab initio* but lacked a conserved repeat element (i.e. divergence > 30%) [79], the region was added back into the probe-design sequence file. Probe sequences were selected using PICKY 2.2 software [81]. These probes were used in solution hybridization to capture our target sequences, followed by NGS on an Illumina Genome Analyzer IIx (Appendix C.1).

Genomic sequences from both strands were captured using overlapping oligonucleotide sequence designs covering 342,075 nt among the 7 genes (Figure 4.1). In total, 11,841 oligonucleotides were synthesized from the transcribed strand consisting of the complete, single copy coding, and flanking regions of *ATM* (3,513), *BRCA1* (1,587), *BRCA2* (2,386), *CDH1* (1,867), *CHEK2* (889), *PALB2* (811), and *TP53* (788). Additionally, 11,828 antisense strand oligos were synthesized (3,497 *ATM*, 1,591 *BRCA1*, 2,395 *BRCA2*, 1,860 *CDH1*, 883 *CHEK2*, 826 *PALB2*, and 776 *TP53*). Any intronic or intergenic regions without probe coverage are most likely due to the presence of conserved repetitive elements or other paralogous sequences.

For regions lacking probe coverage (of ≥ 10 nt, N = 141; 8 in *ATM*, 26 in *BRCA1*, 10 in *BRCA2*, 29 in *CDH1*, 36 in *CHEK2*, 15 in *PALB2*, and 17 in *TP53*), probes were selected based on predicted T_ms similar to other probes, limited alignment to other sequences in the transcriptome (< 10 times), and avoidance of stable, base-paired 2° structures (with unaFOLD) [82, 83]. The average coverage of these sequenced regions was 14.1-24.9% lower than other probe sets, indicating that capture was less efficient, though still successful.

4.2.2 HBOC Samples for Oligo Capture and High-Throughput Sequencing

Genomic DNA from 102 patients previously tested for inherited breast/ovarian cancer without evidence of a predisposing genetic mutation, was obtained from the Molecular Genetics Laboratory (MGL) at the London Health Sciences Centre in London, Ontario,

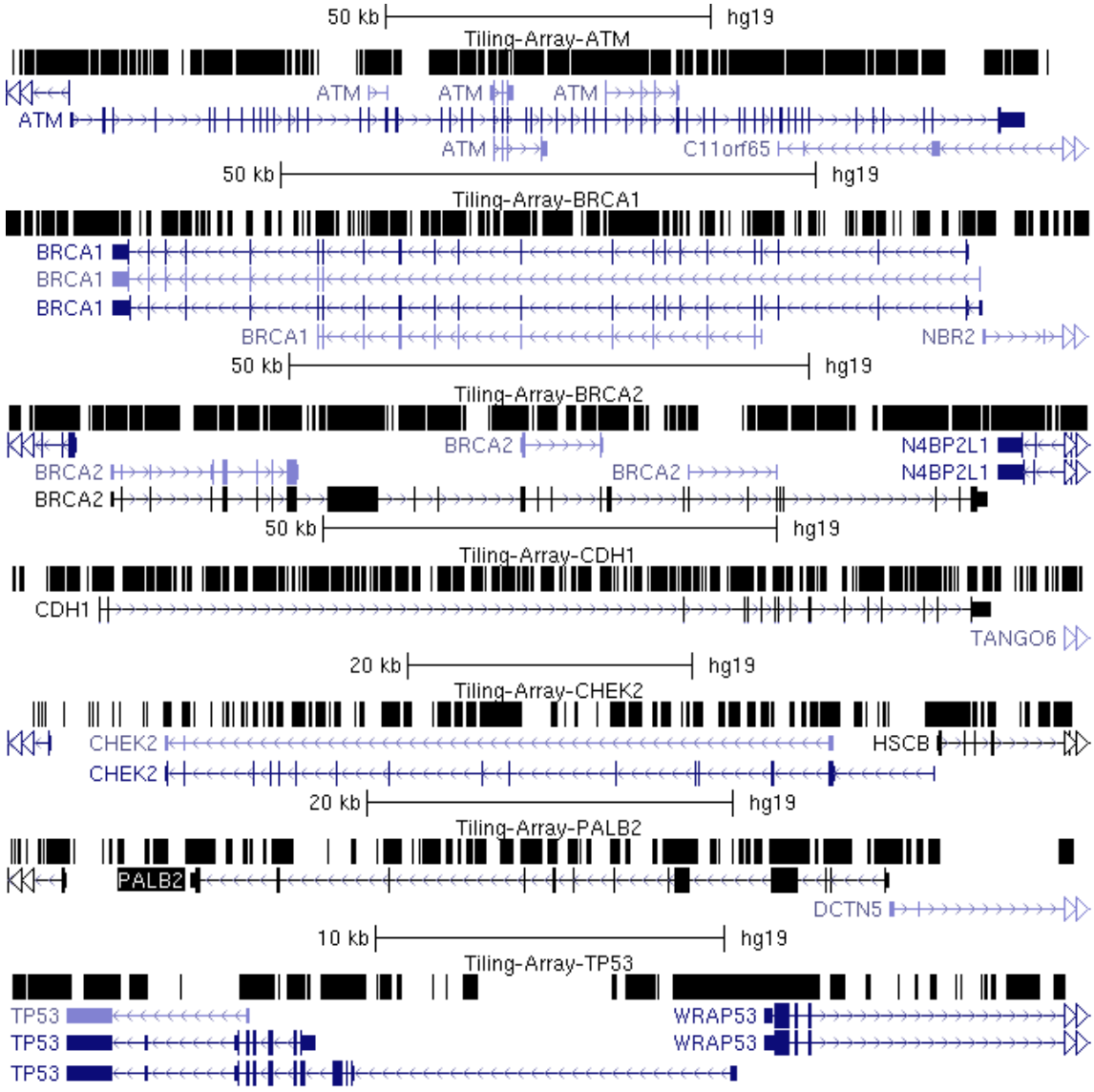

Figure 4.1: Capture probe coverage over sequenced genes. The genomic structure of the 7 genes chosen are displayed with the UCSC Genome Browser. Top row for each gene is a custom track with the "dense" visualization modality selected with black regions indicating the intervals covered by oligonucleotide capture reagent. Regions without probe coverage contain conserved repetitive sequences or correspond to paralogous sequences that are unsuitable for probe design.

Canada. Patients qualified for genetic susceptibility testing as determined by the Ontario Ministry of Health and Long-Term Care *BRCA1* and *BRCA2* genetic testing criteria [84] (see Appendix C.2). The University of Western Ontario research ethics board (REB) approved this anonymized study of these individuals to evaluate the analytical methods presented here. *BRCA1* and *BRCA2* were previously analyzed by Protein Truncation Test (PTT) and Multiplex Ligation-dependent Probe Amplification (MLPA). The exons of several patients (N = 14) had also been Sanger sequenced. No pathogenic sequence change was found in any of these individuals. In addition, one patient with a known pathogenic *BRCA* variant was re-sequenced by NGS as a positive control.

4.2.3 Sequence Alignment and Variant Calling

Variant analysis involved the steps of detection, filtering, IT-based and coding sequence analysis, and prioritization (Figure 4.2). Sequencing data were demultiplexed and aligned to the specific chromosomes of our sequenced genes (hg19) using both CASAVA (Consensus Assessment of Sequencing and Variation; v1.8.2) [85] and CRAC (Complex Reads Analysis and Classification; v1.3.0) [86] software. Alignments were prepared for variant calling using Picard [87] and variant calling was performed on both versions of the aligned sequences using the UnifiedGenotyper tool in the Genome Analysis Toolkit (GATK) [88]. We used the recommended minimum phred base quality score of 30, and results were exported in variant call format (VCF; v4.1). A software program was developed to exclude variants called outside of targeted capture regions and those with quality scores < 50. Variants flagged by bioinformatic analysis (described below) were also assessed by manually inspecting the reads in the region using the Integrative Genomics Viewer (IGV; version 2.3) [89, 90] to note and eliminate obvious false positives (i.e. variant called due to polyhomonucleotide run dephasing, or PCR duplicates that were not eliminated by Picard). Finally, common variants (≥ 1% allele frequency based on dbSNP 142 or > 5 individuals in our study cohort) were not prioritized.

4.2.4 IT-Based Variant Analysis

All variants were analyzed using the Shannon Human Splicing Mutation Pipeline, a genome-scale variant analysis program that predicts the effects of variants on mRNA

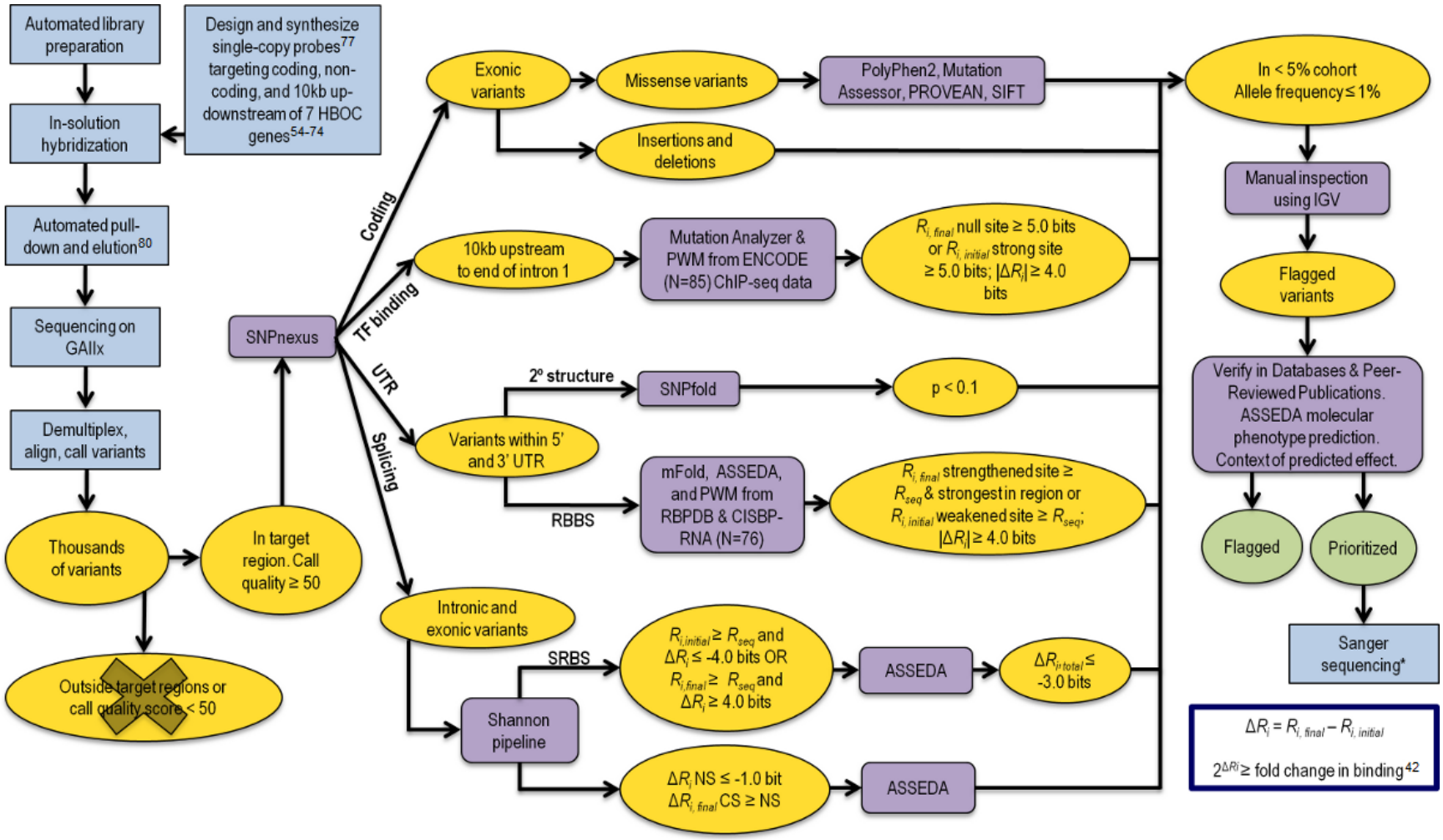

Figure 4.2: Framework for the identification of potentially pathogenic variants. Integrated laboratory processing and bioinformatic analysis procedures for comprehensive complete gene variant determination and analysis. Intermediate datasets

resulting from filtering are represented in yellow and final datasets in green. Non-bioinformatic steps, such as sample preparation are represented in blue and prediction programs in purple. Sequencing analysis yields base calls for all samples. CASAVA [85] and CRAC [86] were used to align these sequencing results to hg19. GATK [88] was used to call variants from this data against GRCh37 release of the reference human genome. Variants with a quality score < 50 and/or call confidence score < 30 were eliminated along with variants falling outside of our target regions. SNPnexus [112–114] was used to identify the genomic location of the variants. Nonsense and indels were noted and prediction tools were used to assess the potential pathogenicity of missense variants. The Shannon Pipeline [91] evaluated the effect of a variant on natural and cryptic SSs, as well as SRFBSs. ASSEDA [38] was used to predict the potential isoforms as a result of these variants. PWMs for 83 TFs were built using an information weight matrix generator based on Bipad [106]. Mutation Analyzer evaluated the effect of variants found 10 kb upstream up to the first intron on protein binding. Bit thresholds ($\boldsymbol{R_i}$ values) for filtering variants on software program outputs are indicated. Variants falling within the UTR sequences were assessed using SNPfold [20], and the most probable variants that alter mRNA structure ($p < 0.1$) were then processed using mFold to predict the effect on stability [83]. All UTR variants were scanned with a modified version of the Shannon Pipeline, which uses PWMs computed from nucleotide frequencies for 28 RBPs in RBPDB [109] and 76 RBPs in CISBP-RNA [110]. All variants meeting these filtering criteria were verified with IGV [89, 90]. * Sanger sequencing was only performed for protein truncating, splicing, and selected missense variants

splicing [91, 92]. Variants were flagged based on criteria reported in Shirley et al. (2013): weakened natural site ≥ 1.0 bits, or strengthened cryptic site (within 300 nt of the nearest exon) where cryptic site strength is equivalent or greater than the nearest natural site of the same phase [91]. The effects of flagged variants were further analyzed in detail using the Automated Splice Site and Exon Definition Analysis (ASSEDA) server [38].

Exonic variants and those found within 500 nt of an exon were assessed for their effects, if any, on SRFBSs [38]. Sequence logos for splicing regulatory factors (SRFs) (SRSF1, SRSF2, SRSF5, SRSF6, hnRNPH, hnRNPA1, ELAVL1, TIA1, and PTB) and their $R_{sequence}$ values (the mean information content [93]) are provided in Caminsky et al. (2015) [36]. Because these motifs occur frequently in unspliced transcripts, only variants with large information changes were flagged, notably those with a) ≥ 4.0 bit decrease, i.e. at least a 16-fold reduction in binding site affinity, with $R_{i,initial} \geq R_{sequence}$ for the particular factor analyzed, or b) ≥ 4.0 bit increase in a site where $R_{i,final} \geq 0$ bits. ASSEDA was used to calculate $R_{i,total}$, with the option selected to include the given SRF in the calculation. Variants decreasing $R_{i,total}$ by < 3.0 bits (i.e. 8-fold) were predicted to potentially have benign effects on expression, and were not considered further.

Activation of pseudoexons through creating/strengthening of an intronic cryptic splice site was also assessed [94]. Changes in intronic cryptic sites, where ΔR_i > 1 bit and $R_{i,final} \geq (R_{sequence}$ − 1 standard deviation [S.D.] of $R_{sequence})$, were identified. A pseudoexon was predicted if a pre-existing cryptic site of opposite polarity (with R_i > [$R_{sequence}$ - 1 S.D.]) and in the proper orientation for formation of exons between 10-250 nt in length was present. In addition, the minimum intronic distance between the pseudoexon and either adjacent natural exon was 100 nt. The acceptor site of the pseudoexon was also required to have a strong hnRNPA1 site located within 10 nt ($R_i \geq R_{sequence}$) [38] to ensure accurate proofreading of the exon [37].

Next, variants affecting the strength of SRFs were analyzed by a contextual exon definition analysis of $\Delta R_{i,total}$. The context refers to the documented splicing activity of an SRF. For example, TIA1 has been shown to be an intronic enhancer of exon definition, so only intronic sites were considered. Similarly, hnRNPA1 proofreads the 3'

SS (acceptor) and inhibits exon recognition elsewhere [95]. Variants that lead to redundant SRFBS changes (i.e. one site is abolished and another proximate site [≤ 2 nt] of equivalent strength is activated) were assumed to have a neutral effect on splicing. If the strength of a site bound by PTB (polypyrimidine tract binding protein) was affected, its impact on binding by other factors was analyzed, as PTB impedes binding of other factors with overlapping recognition sites, but does not directly enhance or inhibit splicing itself [96].

To determine effects of variants on transcription factor (TF) binding, we first established which TFs bound to the sequenced regions of the gene promoters (and first exons) in this study by using ChIP-seq data from 125 cell types (Appendix C.1) [97]. We identified 141 TFs with evidence for binding to the promoters of the genes we sequenced, including c-Myc, C/EBPβ, and Sp1, shown to transcriptionally regulate *BRCA1*, *TP53*, and *ATM*, respectively [98–100]. Furthermore, polymorphisms in *TCF7L2*, known to bind enhancer regions of a wide variety of genes in a tissue-specific manner [101], have been shown to increase risk of sporadic [102] and hereditary breast [103], as well as other types of cancer [104, 105].

IT-based models of the 141 TFs of interest were derived by entropy minimization of the DNase accessible ChIP-seq subsets [106]. Details are provided in Lu R, Mucaki E, and Rogan PK (BioRxiv; http://dx.doi.org/10.1101/042853). While some data sets would only yield noise or co-factor motifs (i.e. co-factors that bind via tethering, or histone modifying proteins [107]), techniques such as motif masking and increasing the number of Monte Carlo cycles yielded models for 83 TFs resembling each factor's published motif. Appendix C.3 contains the final list of TFs and the models we built (described below) [108].

These TFBS models (N = 83) were used to scan all variants called in the promoter regions (10 kb upstream of transcriptional start site to the end of IVS1) of HBOC genes for changes in R_i. Binding site changes that weaken interactions with the corresponding TF (to $R_i \leq R_{sequence}$) are likely to affect regulation of the adjacent target gene. Stringent criteria were used to prioritize the most likely variants and thus only changes to strong

TFBSs ($R_{i,initial} \geq R_{sequence}$), where reduction in strength was significant ($\Delta R_i \geq 4.0$ bits), were considered. Alternatively, novel or strengthened TFBSs were also considered sources of dysregulated transcription. These sites were defined as having$R_{i,final} \geq R_{sequence}$ and as being the strongest predicted site in the corresponding genomic interval (i.e. exceeding the R_i values of adjacent sites unaltered by the variant). Variants were not prioritized if the TF was known to a) enhance transcription and IT analysis predicted stronger binding, or b) repress transcription and IT analysis predicted weaker binding.

Two complementary strategies were used to assess the possible impact of variants within UTRs. First, SNPfold software was used to assess the effect of a variant on 2° structure of the UTR (Appendix C.1) [20]. Variants flagged by SNPfold with the highest probability of altering stable 2° structures in mRNA (where *p*-value < 0.1) were prioritized. To evaluate these predictions, oligonucleotides containing complete wild-type and variant UTR sequences (Appendix C.4) were transcribed *in vitro* and followed by SHAPE analysis, a method that can confirm structural changes in mRNA [44].

Second, the effects of variants on the strength of RBBSs were predicted. Frequency-based, position weight matrices (PWMs) for 156 RNA-binding proteins (RBPs) were obtained from the RNA-Binding Protein DataBase (RBPDB) [109] and the Catalog of Inferred Sequence Binding Preferences of RNA binding proteins (CISBP-RNA) [110, 111]. These were used to compute information weight matrices (based on the method described by Schneider et al. 1984; $N = 147$) (see Appendix C.1) [40]. All UTR variants were assessed using a modified version of the Shannon Pipeline [91] containing the RBPDB and CISBP-RNA models. Results were filtered to include a) variants with $|\Delta R_i| \geq 4.0$ bits, b) variants creating or strengthening sites ($R_{i,final} \geq R_{sequence}$ and the $R_{i,initial} < R_{sequence}$), and c) RBBSs not overlapping or occurring within 10 nt of a stronger, pre-existing site of another RBP.

4.2.5 Exonic Protein-Altering Variant Analysis

The predicted effects of all coding variants were assessed with SNPnexus [112–114], an annotation tool that can be applied to known and novel variants using up-to-date dbSNP and UCSC human genome annotations. Variants predicted to cause premature protein

truncation were given higher priority than those resulting in missense (or synonymous) coding changes. Missense variants were first cross referenced with dbSNP 142 [115]. Population frequencies from the Exome Variant Server [116] and 1000Genomes [117] are also provided. The predicted effects on protein conservation and function of the remaining variants were evaluated by *in silico* tools: PolyPhen-2 [118], Mutation Assessor (release 2) [119, 120], and PROVEAN (v1.1.3) [121, 122]. Default settings were applied and in the case of PROVEAN, the "PROVEAN Human Genome Variants Tool" was used, which includes SIFT predictions as a part of its output. Variants predicted by all four programs to be benign were less likely to have a deleterious impact on protein activity; however this did not exclude them from mRNA splicing analysis (described above in *IT-Based Variant Analysis*). All rare and novel variants were cross-referenced with general mutation databases (ClinVar [123, 124], Human Gene Mutation Database [HGMD] [125, 126], Leiden Open Variant Database [LOVD] [127–134], Domain Mapping of Disease Mutations [DM^2] [135], Expert Protein Analysis System [ExPASy] [136] and UniProt [137, 138]), and gene-specific databases (*BRCA1/2*: the Breast Cancer Information Core database [BIC] [139] and Evidence-based Network for the Interpretation of Germline Mutant Alleles [ENIGMA] [140]; *TP53*: International Agency for Research on Cancer [IARC] [141]), as well as published reports to prioritize them for further workup.

4.2.6 Variant Classification

Flagged variants were prioritized if they were likely to encode a dysfunctional protein (indels, nonsense codon > 50 amino acids from the C-terminus, or abolition of a natural SS resulting in out-of-frame exon skipping) or if they exceeded established thresholds for fold changes in binding affinity based on IT (see *Methods* above). In several instances, our classification was superseded by previous functional or pedigree analyses (reported in published literature or databases) that categorized these variants as pathogenic or benign.

4.2.7 Positive control

We identified the *BRCA1* exon 17 nonsense variant c.5136G>A (chr17:41215907C>T; rs80357418; 2-5A) [142] in the sample that was provided as a positive control. This was

the same mutation identified by the MGL as pathogenic for this patient. We also prioritized another variant in this patient (Table 4.1) [143].

4.2.8 Variant Validation

Protein-truncating, prioritized splicing, and selected prioritized missense variants were verified by Sanger sequencing. Primers of PCR amplicons are indicated in Appendix C.5.

4.2.9 Deletion Analysis

4.2.9.1 Junctional Read Detection

Potential large rearrangements were detected with BreakDancer software [144], which identifies novel genomic rearrangements based on the respective orientation and distance between ends of the same read (and exceeding the lengths of NGS library inserts. This approach can, in theory, approximately localize deletions, duplications, or other types of breakpoints within exons, introns, and regulatory regions (eg. promoters) that could affect gene expression and function. We required at least 4 reads per suspected rearrangement in a sample separated by >700 nt, with each end mapping to proximate genomic reference coordinates to infer a potential deletion. Synthetic and cost limitations in the maximum genomic real estate covered by the capture reagent led to a tradeoff between extending the span of captured genomic intervals and higher tiling densities over shorter sequences, ie. exons, to achieve the level of coverage to reliably detect deletions based on read counts alone.

4.2.9.2 Prioritization based on Potential Hemizygosity

Our complete gene enrichment strategy with independent capture of both genomic strands enabled and facilitated development of a *new* algorithm to identify potential hemizygous genomic intervals in these individuals. In each subject, we first searched for contiguous long stretches (usually >> 1 kb) of non-polymorphic segments with diminished repetitive element content (<10%), which is consistent with the possibility of these regions harboring a deletion. Then, we determined the likelihood of homo- or hemizygosity by comparing the degree of heterozygosity of variants in each of these intervals in for an individual with all of the other individuals sequenced with this protocol

Table 4.1: Prioritized variants in the positive control

Gene	mRNA Protein	rsID (dbSNP 142) Allele Frequency (%)†	Category	Consequence	Ref
BRCA1	c.5136G>A p.Trp1712Ter	rs80357418	Nonsense	151 AA short	[142]
BRCA2	c.3218A>G p.Gln1073Arg	rs80358566	SRFBS	Repressor action of hnRNPA1 at this site abolished (5.2 to 0.4 bits). Blocking action of PTB removed as site is abolished (5.5 to -7.5 bits) and may uncover binding sites of other SRFs.	[143]
			Missense	Listed in ClinVar as conflicting interpretations (likely benign, unknown) and in BIC as unknown clinical importance. 2 *in silico* programs called deleterious. The variant occurs between repeat motifs BRC1 and BRC2 of *BRCA2*, a region in which pathogenic missense mutations have not yet been identified.	
			SRFBS	Repressor action of hnRNPA1 at this site abolished (5.2 to 0.4 bits). Blocking action of PTB removed as site is abolished (5.5 to -7.5 bits) and may uncover binding sites of other SRFs.	

in this population. Regions containing haplotype blocks in strong linkage disequilibrium (LD; from HapMap [145]) were then excluded as candidate deletion intervals. Some individuals without a deletion are expected to be non-polymorphic, because detection of heterozygosity depends on genomic length of the region, marker informativeness, and the level of LD for those markers. We required that > 80% of the control individuals be heterozygous for at least two well-distributed loci within these intervals. Highly informative SNPs with a random genomic distribution in the controls (and other public databases) and which were non-polymorphic in the individual with the suspected deletion were weighted more heavily in inferring potential hemizygosity. This analysis was implemented using a Perl script that identified the most likely intervals of hemizygosity, which were then crossreferenced with the corresponding genomic intervals in HapMap.

4.3 Results

4.3.1 Capture, Sequencing, and Alignment

The average coverage of capture region per individual was 90.8x (range of 53.8 to 118.2x between 32 samples) with 98.8% of the probe-covered nucleotides having ≥ 10 reads. Samples with fewer than 10 reads per nucleotide were re-sequenced and the results of both runs were combined. The combined coverage of these samples was, on average, 48.2x (± 36.2).

The consistency of both library preparation and capture protocols was improved from initial runs, which significantly impacted sequence coverage (Appendix C.1). Of the 102 patients tested, 14 had been previously Sanger sequenced for *BRCA1* and *BRCA2* exons. Confirmation of previously discovered SNVs served to assess the methodological improvements introduced during NGS and ultimately, to increase confidence in variant calling. Initially, only 15 of 49 SNVs in 3 samples were detected. The detection rate of SNVs was improved to 100% as the protocol progressed. All known SNVs (N = 157) were called in subsequent sequencing runs where purification steps were replaced with solid phase reversible immobilization beads and where RNA bait was transcribed the same day as capture. To minimize false positive variant calls, sequence read data were

aligned with CASAVA and CRAC, variants were called for each alignment with GATK, and discrepancies were then resolved by manual review.

GATK called 14,164 unique SNVs and 1,147 indels. Only 3777 (15.3%) SNVs were present in both CASAVA and CRAC-alignments for at least one patient, and even fewer indel calls were concordant between both methods (N = 110; 6.2%). For all other SNVs and indels, CASAVA called 6871 and 1566, respectively, whereas CRAC called 13,958 and 110, respectively. Some variants were counted more than once if they were called by different alignment programs in two or more patients. Intronic and intergenic variants proximate to low complexity sequences tend to generate false positive variants due to ambiguous alignment, a well known technical issue in short read sequence analysis [146, 147], contributing to this discrepancy. For example, CRAC correctly called a 19 nt deletion of *BRCA1* (rs80359876; also confirmed by Sanger sequencing) but CASAVA flagged the deleted segment as a series of false-positives (Appendix C.6). For these reasons, all variants were manually reviewed.

4.3.2 IT-Based Variant Identification and Prioritization

4.3.2.1 Natural SS Variants

The Shannon Pipeline reported 99 unique variants in natural donor or acceptor SSs. After technical and frequency filtering criteria were applied, 12 variants remained (Appendix C.7). IT analysis allowed for the prioritization of 3 variants, summarized in Table 4.2.

First, the novel *ATM* variant c.3747-1G>A (chr11:108,154,953G>A; sample number 7-4F) abolishes the natural acceptor of exon 26 (11.0 to 0.1 bits). ASSEDA reports the presence of a 5.3 bit cryptic acceptor site 13 nt downstream of the natural site, but the effect of the variant on a pre-existing cryptic site is negligible (~0.1 bits). The cryptic exon would lead to exon deletion and frameshift (Figure 4.3A). ASSEDA also predicts skipping of the 246 nt exon, as the $R_{i,final}$ of the natural acceptor is now below $R_{i,minimum}$ (1.6 bits), altering the reading frame. Second, the novel *ATM* c.6347+1G>T (chr11:108188249G>T; 4-1F) abolishes the 10.4 bit natural donor site of exon 44 (ΔR_i= -18.6 bits), and is predicted to cause exon skipping. Finally, the previously reported *CHEK2* variant, c.320-5A>T (chr22:29121360T>A; rs121908700; 4-2B) [148] weakens

Table 4.2: Variants prioritized by IT analysis

Patient ID	Gene	mRNA	rsID (dbSNP 142) Allele Frequency (%)[d]	Information Change			Consequence[f] or Binding Factor Affected
				$R_{i,initial}$ (bits)	$R_{i,initial}$ (bits)	ΔR_i or R_i[e] (bits)	
Abolished Natural SS							
7-4F	*ATM*	c.3747-1G>A[a]	Novel	11.0	0.1	-10.9	Exon skipping and use of alternative splice forms
4-1F	*ATM*	c.6347+1G>T[b]	Novel	10.4	-8.3	-18.6	Exon skipping
Leaky Natural SS							
4-2B	*CHEK2*	c.320-5T>A[a]	rs121908700 0.08	6.8	4.1	-2.7	Leaky splicing with intron inclusion
Activated Cryptic SS							
7-3E	*BRCA1*	c.548-293G>A	rs117281398 0.74	-12.1	2.6	14.7	Cryptic site not expected to be used. Total information for natural exon is stronger than cryptic exon.
7-4A	*BRCA2*	c.7618-269_7618-	Novel	3.9	9.4	5.5	Cryptic site not expected to

		260del10					be used. Total information for natural exon is stronger than cryptic exon.
Pseudoexon formation due to activated acceptor SS							
7-3F	*BRCA2*	c.8332-805G>A	Novel	-9.3	5.4	5.6[e]	6,065/211/592[f]
7-3D	*CDH1*	c.164-2023A>G	rs184740925 0.3	-6.6	4.3	6.5 [e]	61,236/224/1,798[f]
5-3H	*CDH1*	c.2296-174T>A	rs565488866 0.02	7.3	8.5	5.0[e]	1,175/50/124[f]
Pseudoexon formation due to activated donor SS							
3-6A	*BRCA1*	c.212+253G>A	rs189352191 0.08	4.1	6.7	5.2[e]	186/63/1,250[f]
5-2G	*BRCA2*	c.7007+2691G>A	rs367890577 0.02	4.7	7.2	7.7[e]	2,589/103/5,272[f]
Affected TFBSs							
7-4B	*BRCA1*	c.-8895G>A	Novel	10.9	-0.2	-11.1	GATA-3 (*GATA3*)

5-3E 7-4E	*CDH1*	c.-54G>C	rs5030874 0.16	1.7	12.0	10.4	E2F-4 (*E2F4*)
5-2B	*PALB2*	c.-291C>G	rs552824227 0.1	12.1	-1.3	-13.4	GABPα (*GABPA*)
7-2F	*TP53*	c.-28-3132T>C	rs17882863 0.3	-6.3	10.9	17.2	RUNX3 (*RUNX3*)
4-1A	*TP53*	c.-28-1102T>C	rs113451673 0.4	5.1	12.3	7.2	E2F-4 (*E2F4*)
				8.0	12.9	4.8	Sp1 (*SP1*)
Affected RBBSs							
7-4G	*ATM*	c.-244T>A c.-744T>A c.-1929T>A c.-3515T>A	rs539948218 0.04	9.8	-19.9	-29.7	RBFOX
5-3C	*CDH1*	c.*424T>A	Novel	-20.3	9.6	29.9	SF3B4
				8.2	1.8	-6.4	CELF4
7-2E	*CHEK2*	c.-588G>A	rs141568342	10.9	3.7	-7.2	BX511012.1

4-3C.5-4G	*CHEK2*	c.-345C>T[c]	rs137853007	3.3	11.4	8.2	SF3B4
3-1A 4-1H	*TP53*	c.-107T>C c.-188T>C	rs113530090 0.72	10.5	4.5	-6.0	ELAVL1
4-2H 7-2F	*TP53*	c.*1175A>C c.*1376A>C c.*1464A>C	rs78378222 0.26	10.7	4.1	-6.6	KHDRBS1

[a] Confirmed by Sanger sequencing

[b] Ambiguous Sanger sequencing results

[c] Prioritized under missense and was therefore verified with Sanger sequencing. Variant was confirmed

[d] If available

[e] R_i of site of opposite polarity in the pseudoexon

[f] Consequences for pseudoexon formation describe how the intron is divided: “new intron A length/pseudoexon length/new exon B” length.

None of the variants have been previously reported by other groups with the exception of *CHEK2* c.320-5T>A [148].

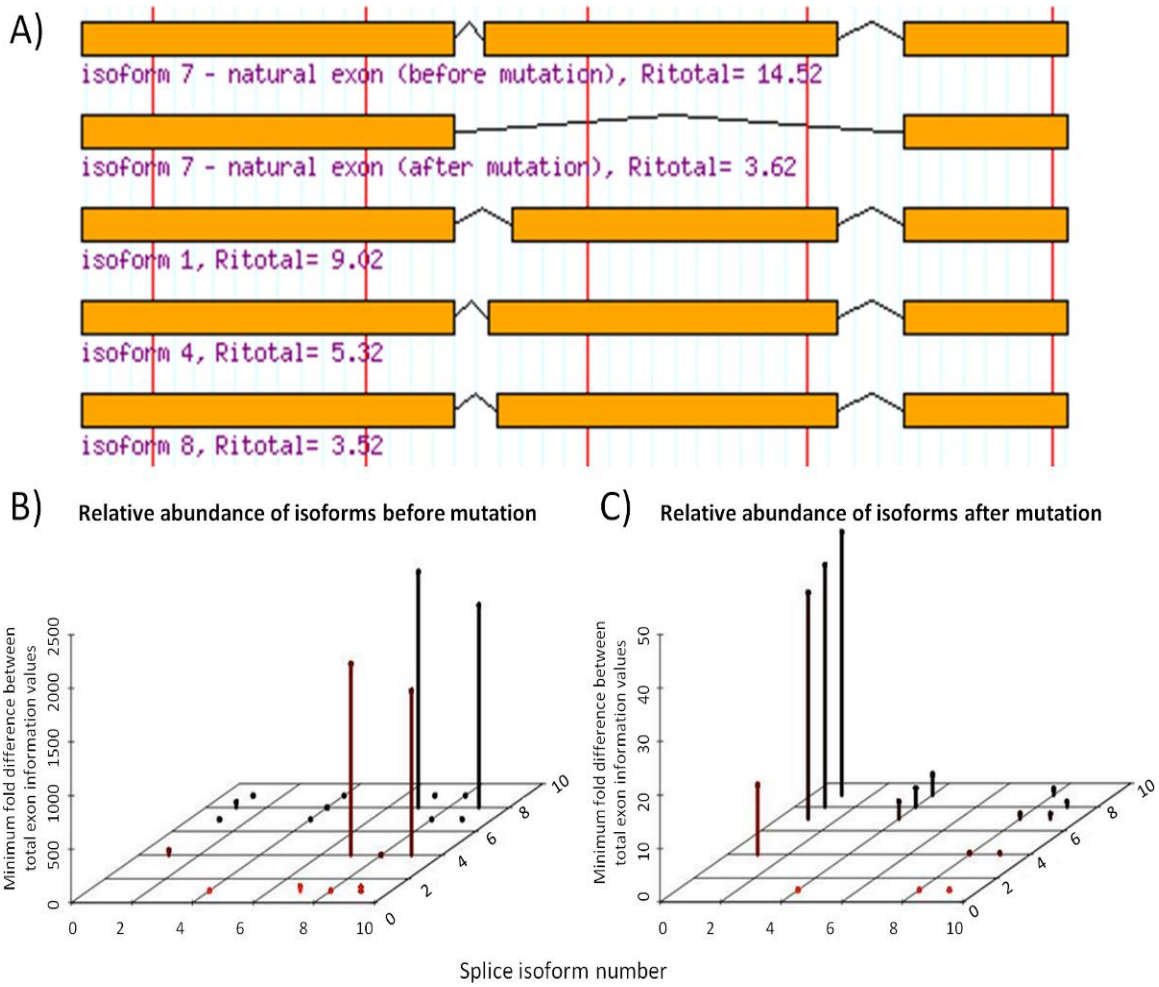

Figure 4.3: Predicted isoforms and relative abundances as a consequence of ATM splice variant c.3747-1G>A. Intronic *ATM* variant c.3747-1G>A abolishes (11.0 to 0.1 bits) the natural acceptor of exon 26 (total of 63 exons). A) ASSEDA predicts skipping of the natural exon ($\boldsymbol{R_{i,initial}}$ from 14.5 to 3.6 bits [an 1910 fold decrease in exon strength]; isoform 7) and/or activation of a preexisting cryptic acceptor site 13 nt downstream ($\boldsymbol{R_{i,total}}$ for cryptic exon = 9.0 bits; isoform 1) of the natural site leading to exon truncation. The reading frame is altered in both mutant isoforms. The other isoforms use weak, alternate acceptor/donor sites leading to cryptic exons with much lower total information. B) Before the mutation, isoform 7 is expected to be the most abundant splice form. C) After the mutation, isoform 1 is predicted to become the most abundant splice form and the wild-type isoform is not expected to be expressed.

the natural acceptor of exon 3 (6.8 to 4.1 bits), and may active a cryptic acceptor (7.4 bits) 92 nt upstream of the natural acceptor site which would shift the reading frame (Figure 4.4). A constitutive, frameshifted alternative isoform of *CHEK2* lacking exons 3 and 4 has been reported, but skipping of exon 3 alone is not normally observed.

Variants either strengthening (N = 4) or slightly weakening (ΔR_i < 1.0 bits; N = 4) a natural site were not prioritized. In addition, we rejected the *ATM* variant (c.1066-6T>G; chr11:108119654T>G; 4-1E and 7-2B), which slightly weakens the natural acceptor of exon 9 (11.0 to 8.1 bits). Although other studies have shown leaky expression as a result of this variant [149], a more recent meta-analysis concluded that this variant is not associated with increased breast cancer risk [150].

4.3.3 Cryptic SS Activation

Two variants produced information changes that could potentially impact cryptic splicing, but were not prioritized for the following reasons (Table 4.2). The first variant, novel *BRCA2* deletion c.7618-269_7618-260del10 (chr13:32931610_32931619del10; 7-4A) strengthens a cryptic acceptor site 245 nt upstream from the natural acceptor of exon 16 ($R_{i,final}$ = 9.4 bits, ΔR_i = 5.5 bits). Being 5.7-fold stronger than the natural site (6.9 bits), two potential cryptic isoforms were predicted, however the exon strengths of both are weaker than the unaffected natural exon ($R_{i,total}$ = 6.6 bits) and thus neither were prioritized. The larger gap surprisal penalties explain the differences in exon strength. The natural donor SS may still be used in conjunction with the abovementioned cryptic SS, resulting in an exon with $R_{i,total}$ = 3.5 bits. Alternatively, the cryptic site and a weak donor site 180 nt upstream of the natural donor (R_i = 0.7 vs 1.4, cryptic and natural donors, respectively) result in an exon with $R_{i,total}$ = 6.5 bits. The second variant, *BRCA1* c.548-293G>A (chr17:41249599C>T; 7-3E), creates a weak cryptic acceptor ($R_{i,final}$ = 2.6 bits, ΔR_i = 6.2 bits) 291 nt upstream of the natural acceptor for exon 8 (ΔR_i = 0.5). Although the cryptic exon is strengthened (final $R_{i,total}$ = 6.9 bits, ΔR_i = 14.7 bits), ASSEDA predicts the level of expression of this exon to be negligible, as it is weaker than the natural exon $R_{i,total}$ = 8.4 bits) due to the increased length of the predicted exon (+291 nt) [38].

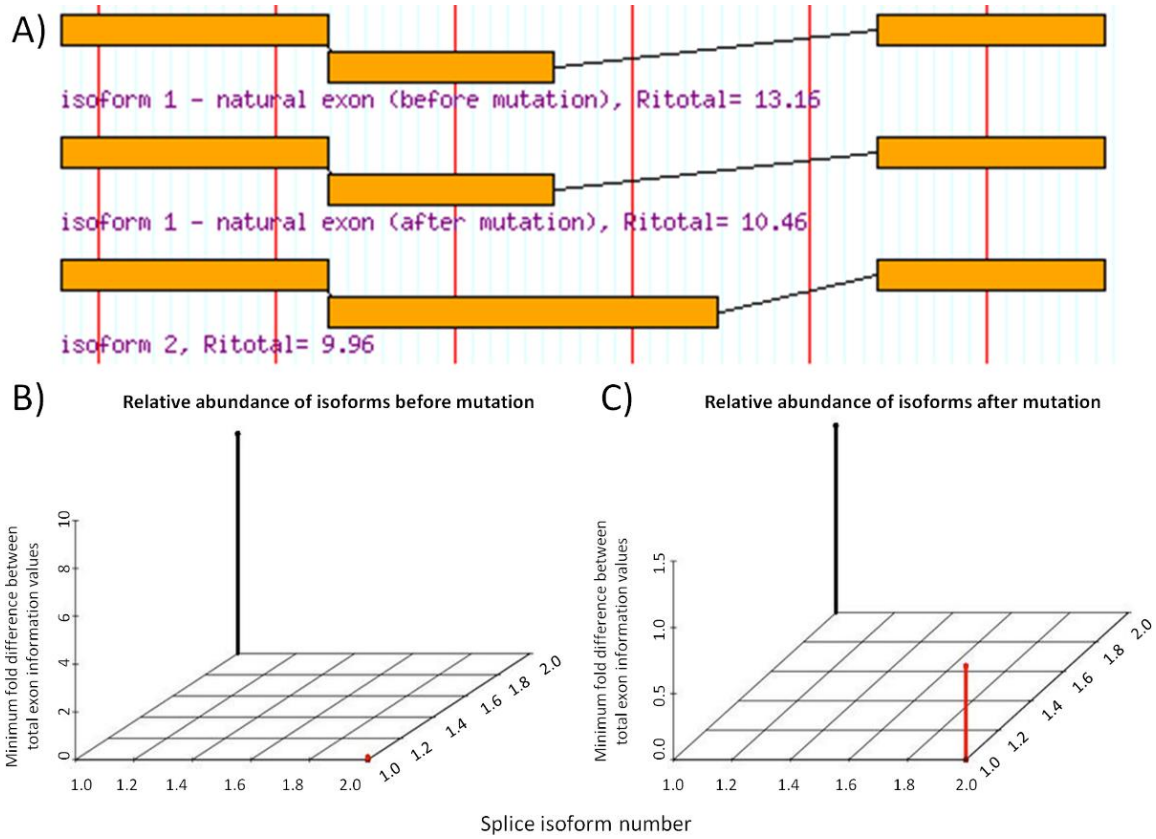

Figure 4.4: Predicted isoforms and relative abundances as a consequence of CHEK2 splice variant c.320-5T>A. Intronic *CHEK2* variant c.320-5T>A weakens (6.8 to 4.1 bits) the natural acceptor of exon 3 (total of 15 exons). A) ASSEDA reports the weakening of the natural exon strength ($\boldsymbol{R_{i,total}}$ reduced from 13.2 to 10.5 bits), which would result in reduced splicing of the exon otherwise known as leaky splicing. A pre-existing cryptic acceptor exists 92 nt upstream of the natural site, leading to a cryptic exon with similar strength to the mutated exon ($\boldsymbol{R_{i,total}}$ = 10.0 bits). This cryptic exon would contain 92 nt of the intron. B) Before the mutation, isoform 1 is expected to be the only isoform expressed. C) After the mutation, isoform 1 (wild-type) is predicted to become relatively less abundant and isoform 2 is expected to be expressed, although less abundant in relation to isoform 1.

4.3.4 Pseudoexon Formation

The Shannon Pipeline initially reported 1583 unique variants creating or strengthening intronic cryptic sites. We prioritized 5 variants, 1 of which is novel (*BRCA2* c.8332-805G>A; 7-3F), that were within 250 nt of a pre-existing complementary cryptic site and have an hnRNPA1 site within 5 nt of the acceptor (Table 4.2). If used, 3 of these pseudoexons would lead to a frameshifted transcript.

4.3.5 SRF Binding

Variants within 500 nt of an exon junction and all exonic variants (N = 4015) were investigated for their potential effects on affinity of sites to corresponding SRFs [38]. IT analysis flagged 54 variants significantly altering the strength of at least one binding site (Appendix C.8). A careful review of the variants, the factor affected, and the position of the binding site relative to the natural SS, prioritized 36 variants (21 novel), of which 4 are in exons and 32 are in introns. As an example, a novel *CHEK2* exon 2 variant c.69C>A (p.Gly23=) is predicted to increase the strength of an hnRNP A1 site (0.7 to 5.3 bits) and decrease total exon strength ($\Delta R_{i,total}$ = -5.7 bits). A similar type of exonic variant in *FANCM*, which was predicted to create exonic hnRNP A1 site by IT, has been shown to bind this exonic repressor and induce exon skipping [37].

4.3.6 TF Binding

We assessed SNVs with models of 83 TFs experimentally shown to bind (Appendix C.3) upstream or within the first exon and intron of our sequenced genes (N = 2177). Thirteen variants expected to significantly affect TF binding were flagged (Appendix C.9). The final filtering step considered the known function of the TF in transcription, resulting in 5 prioritized variants (Table 4.2) in 6 patients (one variant was identified in two patients). Four of these variants have been previously reported (rs5030874, rs552824227, rs17882863, rs113451673) and one is novel (c.-8895G>A; 7-4B).

4.3.7 UTR Structure and Protein Binding

There were 364 unique UTR variants found by sequencing. These variants were evaluated for their effects on mRNA 2° structure (including that of splice forms with

alternate UTRs in the cases of *BRCA1* and *TP53*) through SNPfold, resulting in 5 flagged variants (Table 4.3), all of which have been previously reported.

Analysis of three variants using mFOLD [83] revealed likely changes to the UTR structure (Figure 4.5). Two variants with possible 2° structure effects were common (*BRCA2* c.-52A>G [*N* = 26 samples] and c.*532A>G [*N* = 40]) and not prioritized. The 5'UTR *CDH1* variant c.-71C>G (chr16:68771248C>G; rs34033771; 7-4C) disrupts a double-stranded hairpin region to create a larger loop structure, thus increasing binding accessibility (Figure 4.5A and B). Analysis using RBPDB and CISBP-RNA-derived IT models suggests this variant affects binding by NCL (Nucleolin, a transcription coactivator) by decreasing binding affinity 14-fold ($R_{i,initial}$= 6.6 bits, ΔR_i = -3.8 bits) (Appendix C.10). This RBP has been shown to bind to the 5' and 3' UTR of p53 mRNA and plays a role in repressing its translation [151].

In addition, the *TP53* variant c.*485G>A (NM_000546.5: chr17:7572442C>T; rs4968187) is found at the 3'UTR and was identified in two patients (4-2E and 5-4A). *In silico* mRNA folding analysis demonstrated this variant disrupts a G/C bond of a loop in the highest ranked potential mRNA structure (Figure 4.5C and D). Also, SHAPE analysis showed a difference in 2° structure between the wild-type and mutant (data not shown). IT analysis with RBBS models indicated that this variant significantly increases the binding affinity of SF3B4 by > 48-fold ($R_{i,final}$ = 11.0 bits, ΔR_i = 5.6 bits) (Appendix C.10). This RBP is one of four subunits comprising the splice factor 3B, which binds upstream of the branch-point sequence in pre-mRNA [152].

The third flagged variant also occurs in the 3'UTR of *TP53* (c.*826G>A; chr17:7,572,101C>T; rs17884306), and was identified in 6 patients (2-1A, 7-1B, 5-2A.7-1D, 7-2B, 7-2F, and 7-4C). It disrupts a potential loop structure, stabilizing a double-stranded hairpin, and possibly making it less accessible (Figure 4.5E and F). Analysis using RBPDB-derived models suggests this variant could affect the binding of both RBFOX2 and SF3B4 (Appendix C.10). A binding site for RBFOX2, which acts as a promoter of alternative splicing by favoring the inclusion of alternative exons [153], is created ($R_{i,final}$ = 9.8 bits; ΔR_i = -6.5 bits). This variant is also expected to

Table 4.3: Variants predicted by SNPfold to affect UTR structure

Class[a]	Patient ID	Gene	mRNA	UTR position	rsID (dbSNP 142) Allele Frequency (%)[d]	Rank[e]	*p*-value
F	In 26 patients	*BRCA2*[b]	c.-52A>G	5' UTR	rs206118 14.86	2/900	0.002
F	In 40 patients	*BRCA2*[b]	c.*532A>G	3' UTR	rs11571836 19.75	239/2700	0.089
P	7-4C	*CDH1*[c]	c.-71C>G	5' UTR	rs34033771 0.56	69/600	0.115
F	4-2E 5-4A	*TP53*[b]	c.*485G>A	3' UTR	rs4968187 5.11	169/4500	0.038
F	2-1A, 7-1B, 5-2A, 7-1D, 7-2B, 7-2F, 7-4C	*TP53*[b]	c.*826G>A	3' UTR	rs17884306 5.71	371/4500	0.082

[a] F:Flagged; P:Prioritized

[b] Long Range UTR SNPfold Analysis

[c] Local Range SNPfold Analysis

[d] If available

[e] Rank of the SNP, in terms of how much it changes the mRNA structure compared to all other possible mutations.

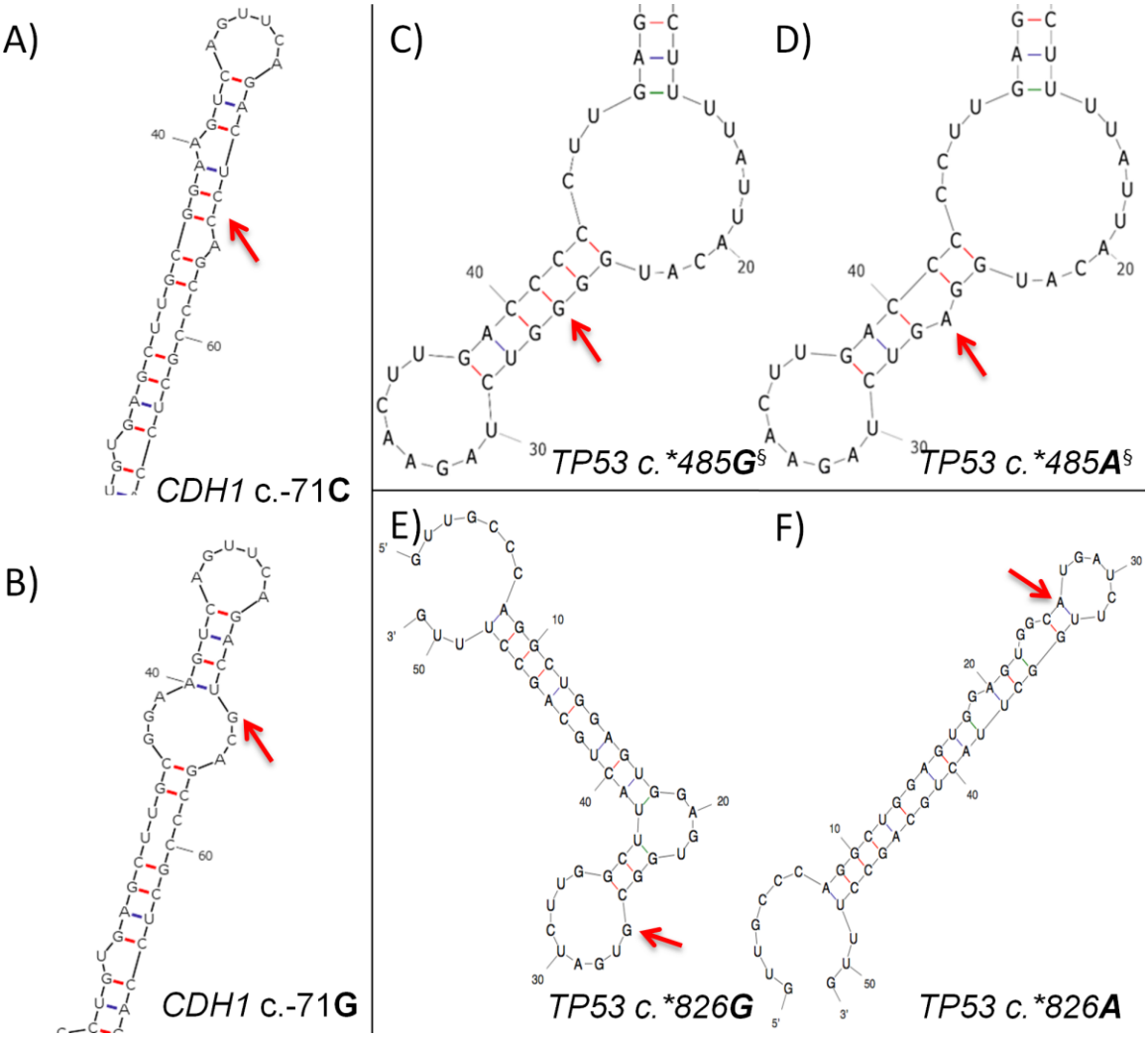

Figure 4.5: Predicted alteration in UTR structure using mFOLD for variants flagged by SNPfold. Wild-type and variant structures are displayed, with the variant indicated by a red arrow. A) Predicted wild-type structure of *CDH1* 5'UTR surrounding c.-71. B) Predicted *CDH1* 5'UTR structure due to c.-71C>G variant. C) Predicted wild-type *TP53* 3'UTR structure surrounding c.*485. D) Predicted *TP53* 5'UTR structure due to c.*485G>A variant. E) Predicted wild-type *TP53* 3'UTR structure surrounding c.*826. F) Predicted *TP53* 5'UTR structure due to c.*826G>A variant. §SHAPE analysis revealed differences in reactivity between mutant and variant mRNAs, confirming alterations to 2° structure.

simultaneously abolish a SF3B4 binding site ($R_{i,final}$ = -20.3 bits; ΔR_i = -29.9 bits). RBPDB- and CISBP-RNA-derived information model analysis of all UTR variants resulted in the prioritization of 1 novel, and 5 previously-reported variants (Table 4.2). No patient within the cohort exhibited more than one prioritized RBBS variant.

To evaluate the background rate of prioritizing variants flagged by this method, all 5' and 3' UTR SNVs in dbSNP144 for the 7 genes sequenced (excluding those already flagged in Table 4.3) were evaluated by SNPfold and our RBP information models. Of 1207 SNVs, only 10 were prioritized with both methods, which results in a background rate of 0.83%.

4.3.8 Exonic Variants Altering Protein Sequence

Exonic variants called by GATK (N = 245) included insertions, deletions, nonsense, missense, and synonymous changes.

4.3.8.1 Protein-Truncating Variants

We identified 3 patients with different indels (Table 4.4). One was a *PALB2* insertion c.1617_1618insTT (chr16:23646249_23646250insAA; 5-3A) in exon 4, previously reported in ClinVar as pathogenic. This mutation results in a frameshift and premature translation termination by 626 residues, abolishing domain interactions with RAD51, BRCA2, and POLH [137]. We also identified two known frameshift mutations in *BRCA1*: c.4964_4982del19 in exon 15 (chr17:41222949_41222967del19; rs80359876; 5-1B) and c.5266_5267insC in exon 19 (chr17:41209079_41209080insG; rs397507247; 5-3C) [148, 154]. Both are indicated as pathogenic and common in the BIC Database due to the loss of one or both C-terminal BRCT repeat domains [137]. Truncation of these domains produces instability and impairs nuclear transcript localization [155], and this bipartite domain is responsible for binding phosphoproteins that are phosphorylated in response to DNA damage [156, 157].

We also identified 4 nonsense mutations, one of which was novel in exon 4 of *PALB2* (c.1042C>T; chr16:23646825G>A; 4-4D). Another in *PALB2* has been previously reported (c.1240C>T; chr16:23646627G>A; rs180177100; 7-3A) [58]. As a

Table 4.4: Variants resulting in premature protein truncation

Patient ID	Gene	Exon	mRNA Protein	rsID (dbSNP 142) Allele Frequency (%)[c]	ClinVar[d,e,f]	Details	Ref
Insertions/Deletions							
5-1B	*BRCA1*	15 of 23	c.4964_4982del19[a] p.Ser1655Tyrfs	rs80359876	6[d]; Pathogenic/likely pathogenic[e]; Familial breast and breast-ovarian cancer, Hereditary cancer-predisposing syndrome[f].	STOP at p.1670 193 AA short	-
5-3C	*BRCA1*	19 of 23	c.5266_5267insC[a] p.Gln1756Profs	rs397507247	13[d]; Pathogenic, risk factor[e]; Familial breast, breast-ovarian, and pancreatic cancer, Hereditary cancer-predisposing syndrome[f].	STOP at p.1788 75 AA short	[148, 154]
5-3A	*PALB2*	4 of 13	c.1617_1618insTT[a] p.Asn540Leufs	-	1[d]; Pathogenic[e]; Hereditary cancer-predisposing syndrome[f].	STOP at p.561 626 AA short	-
Stop Codons							

7-1G	*BRCA2*	15 of 27	c.7558C>T[b] p.Arg2520Ter	rs80358981	5[d]; Pathogenic[e]; Familial breast, and breast-ovarian cancer, Hereditary cancer-predisposing syndrome[f].	899 AA short	[158]
4-4A	*BRCA2*	25 of 27	c.9294C>G[a] p.Tyr3098Ter	rs80359200	3[d]; Pathogenic[e]; Familial breast and breast-ovarian cancer[f].	321 AA short	[159]
7-3A	*PALB2*	4 of 13	c.1240C>T[a] p.Arg414Ter	rs180177100	3[d]; Pathogenic[e]; Familial breast cancer, Hereditary cancer-predisposing syndrome[f].	773 AA short	[58]
4-4D	*PALB2*	4 of 13	c.1042C>T[a] p.Gln348Ter	Novel	-	839 AA short	-

[a] Confirmed by Sanger sequencing

[b] Not confirmed by Sanger sequencing

[c] If available

[d] Number of submissions

[e] Clinical significance

[f] Condition(s)

consequence, functional domains of PALB2 that interact with BRCA1, RAD51, BRCA2, and POLH are lost [137]. Two known nonsense mutations were found in *BRCA2*, c.7558C>T in exon 15 [158] and c.9294C>G in exon 25 [159]. The first (chr13:32930687C>T; rs80358981; 7-1G) causes the loss of the BRCA2 region that binds FANCD2, responsible for loading BRCA2 onto damaged chromatin [160]. The second (chr13:32968863C>G, rs80359200; 4-4A) does not occur within a known functional domain, however the transcript is likely to be degraded by nonsense mediated decay [161].

4.3.8.2 Missense

GATK called 61 missense variants, of which 18 were identified in 6 patients or more and 19 had allele frequencies > 1.0% (Appendix C.11). The 40 remaining variants (15 *ATM*, 8 *BRCA1*, 9 *BRCA2*, 2 *CDH1*, 2 *CHEK2*, 3 *PALB2*, and 1 *TP53*) were assessed using a combination of gene specific databases, published classifications, and 4 *in silico* tools (Appendix C.12). We prioritized 27 variants, 2 of which were novel. None of the non-prioritized variants were predicted to be damaging by more than 2 of 4 conservation-based software programs.

4.3.9 Variant Classification

Initially, 15,311 unique variants were identified by complete gene sequencing of 7 HBOC genes. Of these, 132 were flagged after filtering, and further reduced by IT-based variant analysis and consultation of the published literature to 87 prioritized variants. Figure 4.6 illustrates the decrease in the number of unique variants per patient at each step of our identification and prioritization process. The distribution of prioritized variants by gene is 34 in *ATM*, 13 in *BRCA1*, 11 in *BRCA2*, 8 in *CDH1*, 6 in *CHEK2*, 10 in *PALB2*, and 5 in *TP53* (Appendix C.13), which are categorized by type in Table 4.5.

Three prioritized variants have multiple predicted roles: *ATM* c.1538A>G in missense and SRFBS, *CHEK2* c.190G>A in missense and UTR binding, and *CHEK2* c.433C>T in missense and UTR binding. Of the 102 patients that we sequenced, 72 (70.6%) exhibited at least one prioritized variant, and some patients harbored more than one prioritized

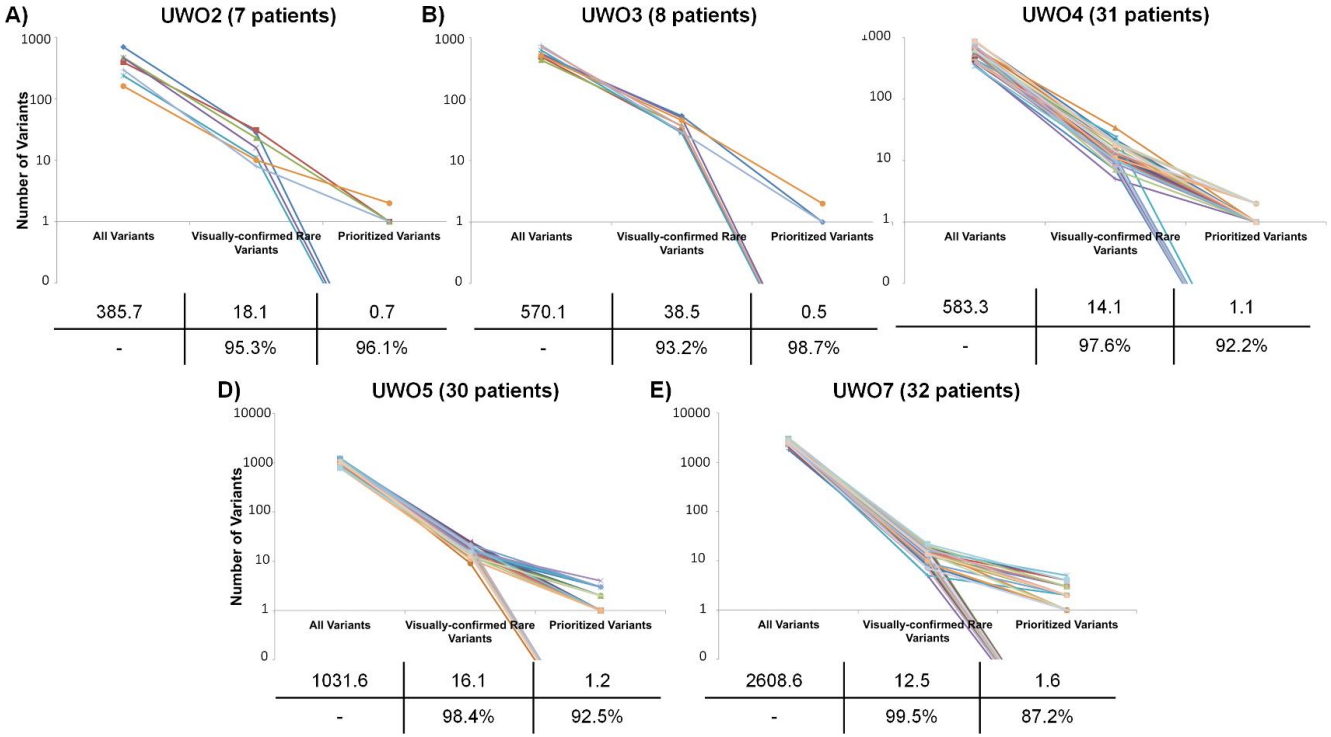

Figure 4.6: Ladder plot representing variant identification and prioritization. Each line is representative of a different sample in each sequencing run (A-E), illustrating the number of unique variants at important steps throughout the variant prioritization process. The left-most point indicates the total number of unique variants. The second point represents the number of unique variants remaining after common (> 5 patients within cohort and/or ≥ 1.0% allele frequency) and false-positive variants were removed. The right-most point represents the final number of unique. No variants were prioritized in the following patients: 2-1A, 2-5A, 2-6A, 3-2A, 3-3A, 3-4A, 3-5A, 3-8A, 4-1B, 4-2C, 4-2F, 4-3B, 4-3D, 4-4B, 4-4E, 5-1G, 5-1H, 5-3D, 5-4C, 5-4D, 5-4F, 5-4G, 5-4H, 7-1B, 7-1C, 7-1D, 7-1H, 7-2B, 7-2C, 7-2H, 7-3H, 7-4A, 7-4D, 7-4H. The average number of variants per patient at each step is indicated in a table below each plot, along with the percent reduction in variants from one step to another.

Table 4.5. Summary of prioritized variants by gene

	Indel	Nonsense	Missense	Natural Splicing	Cryptic Splicing	Pseudoexon	SR Factor	TF	UTR Structure	UTR Binding	Total
ATM	0	0	14	2	0	0	18	0	0	1	34[a]
BRCA1	2	0	2	0	0	1	7	1	0	0	13
BRCA2	0	2	3	0	0	2	4	0	0	0	11
CDH1	0	0	2	0	0	2	1	1	1	1	8
CHEK2	0	0	2	1	0	0	3	0	0	2	6[a]
PALB2	1	2	3	0	0	0	3	1	0	0	10
TP53	0	0	1	0	0	0	0	2	0	2	5
Total	3	4	27	3	0	5	36	5	1	6	

Three variants were prioritized under multiple categories: *ATM* chr11:108121730A>G (missense and SRFBS), *CHEK2* chr22:29121242G>A (missense, UTR binding), and *CHEK2* chr22:29130520C>T (missense, UTR binding).

[a] Counts represent the number of unique variants identified (i.e. a variant is not counted twice if it appeared in multiple individuals).

variant (*N* = 33; 32%). Appendix C.14 presents a summary of all flagged and prioritized variants for patients with at least one prioritized variant.

4.3.10 Prioritization of Potential Deletions

Using BreakDancer, none of the individuals analyzed exhibited large rearrangements that met the level of stringency required, but a small intragenic rearrangement in *BRCA1* was identified and confirmed by Sanger sequencing. Attempts to detect deletions with BreakDancer only flagged single, non-contiguous paired-end reads, rather than a series of reads clustered within the same region within the same individual, which would be necessary to indicate the presence of a true deletion or structural rearrangement.

After prioritizing individuals for potential hemizygosity in the sequenced regions, potential deletions were detected in *BRCA2* and *CDH1*. Patient UWO5-4D exhibited a non-polymorphic 32.1 kb interval in *BRCA2* spanning introns 1 to 13 that was absent from all of the other individuals (chr13:32890227-32922331). Haploview (hapmap.org) showed very low levels of linkage disequilibrium in this region. The potential deletion may extend further downstream, however the presence of a haploblock covering the entire sequenced interval beyond exon 11 with significant LD precludes delineation of the telomeric breakpoint. We also flagged a non-polymorphic 2.6 kb interval near the 3' end of *CDH1* in 6 individuals (UWO3-5, UWO4-2C, UWO4-4E, UWO4-4F, UWO4-2G, UWO5-2H). This is a low LD region spanning chr16:68861286-68863887 that includes exons 14 and 15, and is polymorphic in all of the other individuals sequenced. *CDH1* mutations are characteristically present in families with predisposition to gastric cancer, however breast cancer frequently co-occurs [69]. A study of *CDH1* deletions in inherited gastric cancer identified two families with deletions that overlap the intervals prioritized in the present study [162].

4.3.11 Comparison to Combined Annotation Dependent Depletion

The analysis and prioritization of non-coding variants can also be accomplished using Combined Annotation Dependent Depletion (CADD; [163]), which uses known and simulated variants to compute a C-score, an ad hoc measure of how deleterious is likely to be. The suggested C-score cutoff is between 10-20, though it is stated that any selected

cutoff value would be arbitrary (http://cadd.gs.washington.edu/info). This contrasts with information-based methods, which are based on thermodynamically-defined thresholds. To directly compare methods, CADD scores were obtained for all prioritized or flagged SNVs. Half of prioritized variants met this cutoff (C>10), while only 28.6% of flagged variants did the same. All prioritized nonsense variants (4/4) and 26/27 missense variants had strong C-scores. Prioritized non-coding variant categories that correlated well with CADD include natural splicing variants (4/4), UTR structure variants (1/1), and RBPs (4/6).Weakly correlated variants included those affecting SRFBPs (5/36), TFBS (2/5), and pseudoexon activating variants (0/5). Missense mutations comprised 75% of the flagged variants with C>10. The aforementioned flagged splicing variant *ATM* c.1066-6T>G also exceeded the threshold C value (*C* = 11.9). Meanwhile, the flagged *TP53* variant, shown by SHAPE analysis to alter UTR structure, did not (*C* = 5.3). Despite consistency between some variant categories, the underlying assumptions of each approach probably explain why these results differ for non-coding variants. The limited numbers validated, deleterious non-coding variants also contributes to the accuracy of these predictions [163].

4.3.12 Variant Verification

We verified prioritized protein-truncating (*N* = 7) and splicing (*N* = 4) variants by Sanger sequencing (Table 4.2 and Table 4.4, respectively). In addition, two missense variants (*BRCA2* c.7958T>C and *CHEK2* c.433C>T) were re-sequenced, since they are indicated as likely pathogenic/pathogenic in ClinVar (Appendix C.12). All protein-truncating variants were confirmed, with one exception (*BRCA2* c.7558C>T, no evidence for the variant was present for either strand). Two of the mRNA splicing mutations were confirmed on both strands, while the other two were confirmed on a single strand (*ATM* c.6347+1G>T and *ATM* c.1066-6T>G). Both documented pathogenic missense variants were also confirmed.

4.4 Discussion

NGS technology offers advantages in throughput and variant detection [126], but the task of interpreting the sheer volume of variants in complete gene or genome data can be

daunting. The whole genome of a Yoruban male contained approximately 4.2 million SNVs and 0.4 million structural variants [164]. The variant density in the present study (average 948 variants per patient) was 5.3-fold lower than the same regions in HapMap sample NA12878 in Illumina Platinum Genomes Project (5029 variants) [165]. The difference can be attributed primarily to the exclusion of polymorphisms in highly repetitive regions in our study.

Conventional coding sequence analysis, combined with an IT-based approach for regulatory and splicing-related variants, reduced the set to a manageable number of prioritized variants. Unification of non-coding analysis of diverse protein-nucleic acid interactions using the IT framework accomplishes this by applying thermodynamic-based thresholds to binding affinity changes and by selecting the most significant binding site information changes, regardless of whether the motifs of different factors overlap.

Previously, rule-based systems have been proposed for variant severity classification [166, 167]. Functional validation and risk analyses of these variants are a prerequisite for classification, but this would not be practical to accomplish without first limiting the subset of variants analyzed. With the exception of some (but not all [37]) protein truncating variants, classification is generally not achievable by sequence analysis alone. Only a minority of variants with extreme likelihoods of pathogenic or benign phenotypes are clearly delineated because only these types of variants are considered actionable [166, 167]. The proposed classification systems preferably require functional, co-segregation, and risk analyses to stratify patients. Nevertheless, the majority of variants are VUS, especially in the case of variants occurring beyond exon boundaries. Of the 5713 variants in the BIC database, the clinical significance of 4102 *BRCA1* and *BRCA2* variants are either unknown (1904) or pending (2198), and only 1535 have been classified as pathogenic (Class 5) [168]. Our results cannot be considered equivalent to validation, which usually include expression assays [36] or the use of RNA-seq data [169] (splicing), qRT-PCR [170] (transcription), SHAPE analysis (mRNA 2° structure) [44], or binding assays to determine functional effects of variants. Classification of VUS in *BRCA1* and *BRCA2* by the ENIGMA Consortium addresses mRNA splicing and missense variants. Criteria define risk based on whether the variant occurs within a protein structural

domain, the impact on protein function, and the segregation pattern of variant with disease in pedigrees [171]. These guidelines cannot be fully implemented here for several reasons: a) patients were anonymized in this study, precluding segregation analysis, b) the splicing mutation guideline does not take into account predicted leaky or cryptic splice site mutations, nor other non-canonical changes that have been demonstrated to alter the expression of these and numerous other genes, c) conserved domains have not been identified in regions of the proteins encoded by these genes, especially *BRCA2*, where many missense mutations reside, and d) the guidelines are currently silent as to the potential impact of regulatory variants affecting splicing, RNA stability, and transcriptional regulation.

While miRNA variant prediction program mrSNP [172] was used to evaluate all of the 3' UTR variants, 41.4% of the variants were predicted to alter the stability of the miRNA-target mRNA duplex for at least one miRNA expressed in breast tissue. However, only 2 of these interactions could be confirmed using TarBase [173], and these variants could not be prioritized for disruption of miRNA regulation. Other post-transcriptional processes, including miRNA regulation, that were not addressed in this study, may also be amenable to IT-based modeling. With the proposed approach, functional prediction of variants could precede or at least inform the classification of VUS.

It is unrealistic to expect all variants to be functionally analyzed, just as it may not be feasible to assess family members for a suspected pathogenic variant detected in a proband. The prioritization procedure reduces the chance that significant variants have been overlooked. Capturing coding and non-coding regions of HBOC-related genes, combined with the framework for assessing variants, balances the need to comprehensively detect all variation in a gene panel with the goal of identifying variants likely to be phenotypically relevant.

The location of variants in relation to known protein domains was documented in this study, but was not directly incorporated into our prioritization method. The locations and impact of splicing mutations in *BRCA1* and *BRCA2* were mapped to the known functional domains of the encoded proteins [174]. A high concentration of variants

predicted to result in splicing changes occurred in the BRCT, RING finger, and NLS domains of *BRCA1*. However, *BRCA2* variants were generally concentrated outside of known functional domains (aside from the C-terminal domain). Because of these inconsistencies, domain-mapping was not integrated with IT based prioritization. However, where adequate information on structure-function relationships is available (eg. *TP53*), we suggest that such analysis be carried out subsequent to IT-based variant prioritization.

4.4.1 Non-coding Variants

Although coding variants are typically the sole focus of a molecular diagnostic laboratory (with the exception of the canonical dinucleotide positions within splice sites), non-coding mutations have long been known to be disease causing [19, 36, 175–183]. In this study, variant density in non-coding regions significantly exceeded exonic variants by > 60-fold, which, in absolute terms, constituted 1.6% of the 15,311 variants. This is comparable to whole genome sequencing studies, which typically result in 3-4 million variants per individual, with < 2% occurring in protein coding regions [184]. IT analysis prioritized 3 natural SS, 36 SRFBS, 5 TFBS, and 6 RBBS variants and 5 predicted to create pseudoexons. Two SS variants in *ATM* (c.3747-1G>A and c.6347+1G>T) were predicted to completely abolish the natural site and cause exon skipping. A *CHEK2* variant (c.320-5A>T) was predicted to result in leaky splicing.

The IT-based framework evaluates all variants on a common scale, based on bit values, the universal unit that predicts changes in binding affinity [185]. A variant can alter the strength of one or a "set" of binding sites; the magnitude and direction of these changes is used to rank their significance. The models used to derive information weight matrices take into account the frequency of all observed bases at a given position of a binding motif, making them more accurate than consensus sequence and conservation-based approaches [36].

IT has been widely used to analyze natural and cryptic SSs [36], but its use in SRFBS analysis was only introduced recently [38]. For this reason, we assigned conservative, minimum thresholds for reporting information changes. Although there are examples of

disease-causing variants resulting in small changes in R_i [174, 186–192], the majority of deleterious splicing mutations that have been verified functionally, produce large information changes. Among 698 experimentally deleterious variants in 117 studies, only 1.96% resulted in < 1.0 bit change [36]. For SRFBS variants, the absolute information changes for deleterious variants ranged from 0.2 - 17.1 bits (mean 4.7 ± 3.8). This first application of IT in TFBS and RBBS analysis, however, lacks a large reference set of validated mutations for the distribution of information changes associated with deleterious variants. The release of new ChIP-seq datasets will enable IT models to be derived for TFs currently unmodeled and will improve existing models [193].

Pseudoexon activation results in disease-causing mutations [194], however such consequences are not customarily screened for in mRNA splicing analysis. IT analysis was used to detect variants that predict pseudoexon formation and 5 variants were prioritized. Previously, we have predicted experimentally proven pseudoexons with IT (Ref 42: Table 4.2, No #2; and Ref 195: Table 4.2, No #7) [42, 195]. Although it was not possible to confirm prioritized variants in the current study predicted to activate pseudoexons because of their low allele frequencies, common intronic variants that were predicted to form pseudoexons were analyzed. We then searched for evidence of pseudoexon activation in mapped human EST and mRNA tracks [196] and RNA-seq data of breast normal and tumour tissue from the Cancer Genome Atlas project [15]. One of these variants (rs6005843) appeared to splice the human EST HY160109 [197] at the predicted cryptic splice site and is expressed within the pseudoexon boundaries.

Variants that were common within our population sample (i.e. occurring in > 5 individuals) and/or common in the general population (> 1.0% allele frequency) reduced the list of flagged variants substantially. This is now a commonly accepted approach for reducing candidate disease variants [166], based on the principle that the disease-causing variants occur at lower population frequencies. Variants occurring in > 5 patients all either had allele frequencies above 1.0% or, as shown previously, resulted in very small ΔR_i values [198].

The genomic context of sequence changes can influence the interpretation of a particular variant [36]. For example, variants causing significant information changes may be interpreted as inconsequential if they are functionally redundant or enhancing existing binding site function (see *IT-Based Variant Analysis* for details). Our understanding of the roles and context of these cognate protein factors is incomplete, which affects confidence in interpretation of variants that alter binding. Also, certain factors with important roles in the regulation of these genes, but that do not bind DNA directly or in a sequence-specific manner (eg. CtBP2 [199]), could not be included. Therefore, some variants may have been incorrectly excluded.

4.4.2 Prioritization of Potential Deletions

Although individuals can be prioritized based on potential hemizygosity, this does not definitively identify deletions. Nevertheless, it should be possible to prioritize those individuals worthy of further detailed diagnostic workup. It has not escaped our attention that the weighted probabilities obtained from this analysis could be represented and formalized using the same units of Shannon information (in bits) as the other sequence changes we have described, analogous to single or multinucleotide gene variants predicted to affect nucleic acid binding sites. Full development and validation of this method is in progress.

4.4.3 Coding Sequence Changes

We also identified 4 nonsense and 3 indels in this cohort. In one individual, a 19 nt *BRCA1* deletion in exon 15 causes a frameshift leading to a stop codon within 14 codons downstream. This variant, rs80359876, is considered clinically relevant. Interestingly, this deletion overlaps two other published deletions in this exon (rs397509209 and rs80359884). This raises the question as to whether this region of the *BRCA1* gene is a hotspot for replication errors. DNA folding analysis indicates a possible 15 nt long stem-loop spanning this interval as the most stable predicted structure (data not shown). This 15 nt structure occurs entirely within the rs80359876 and rs397509209 deletions and partially overlaps rs80359884 (13 of 15 nt of the stem loop). It is plausible that the 2°

structure of this sequence predisposes to a replication error that leads to the observed deletion.

Missense coding variants were also assessed using multiple *in silico* tools and evaluated based on allele frequency, literature references, and gene-specific databases. Of the 27 prioritized missense variants, the previously reported *CHEK2* variant c.433G>A (chr22:29121242G>A; rs137853007) stood out, as it was identified in one patient (4-3C.5-4G) and is predicted by all 4 *in silico* tools to have a damaging effect on protein function. Accordingly, Wu et al. (2001) demonstrated reduced *in vitro* kinase activity and phosphorylation by ATM kinase compared to the wild-type CHEK2 protein [200], presumably due to the variant's occurrence within the forkhead homology-associated domain, involved in protein-phosphoprotein interactions [201]. Implicated in Li-Fraumeni syndrome, known to increase the risk of developing several types of cancer including breast [202, 203], the *CHEK2*: c.433G>A variant is expected to result in a misfolded protein that would be targeted for degradation via the ubiquitin-proteosome pathway [204]. Another important missense variant is c.7958T>C (chr13:32,936,812T>C; rs80359022; 4-4C) in exon 17 of *BRCA2*. Although classified as being of unknown clinical importance in both BIC and ClinVar, it has been classified as pathogenic based on posterior probability calculations [205].

It is unlikely that all prioritized variants are pathogenic in patients carrying more than one prioritized variant. Nevertheless, a polygenic model for breast cancer susceptibility, whereby multiple moderate and low-risk alleles contribute to increased risk of HBOC may also account for multiple prioritized variants [206, 207]. There was a significant fraction of patients (29.4%) in whom no variants were prioritized. This could be due to a) the inability of the analysis to predict a variant affecting the binding sites analyzed, b) a pathogenic variant affecting a function that was not analyzed or in a gene that was not sequenced, c) a large rearrangement/deletion where both breakpoints occur beyond the captured genomic intervals (which is unlikely, as this would have been observed as an extended non-polymorphic sequence), or d) the significant family history was not due to heritable, but instead to shared environmental influences.

BRCA coding variants were found in individuals who were previously screened for lesions in these genes, suggesting this NGS protocol is a more sensitive approach for detecting coding changes. However, previous testing of a number of these patients had been predominantly based on PTT and MLPA, which have lower sensitivity for detecting mutations than sequence analysis. Nevertheless, we identified 2 *BRCA1* and 2 *BRCA2* variants predicted to encode prematurely truncated proteins. Fewer non-coding *BRCA* variants were prioritized (15.7%) than expected by linkage analysis [49], however this presumes at least 4 affected breast cancer diagnoses per pedigree, and, in the present study, the number of affected individuals per family was not known.

Prioritization of a variant does not equate with pathogenicity. Some prioritized variants may not increase risk, but may simply modify a primary unrecognized pathogenic mutation. A patient with a known *BRCA1* nonsense variant, used as a positive control, was also found to possess an additional prioritized variant in *BRCA2* (missense variant chr13:32911710A>G), which was flagged by PROVEAN and SIFT as damaging, as well as flagged for changing an SRFBS for abolishing a PTB site (while simultaneously abolishing an exonic hnRNPA1 site). This variant has been identified in cases of early onset prostate cancer and is considered a VUS in ClinVar [143]. Similarly, variants prioritized in multiple patients may act as risk modifiers rather than pathogenic mutations. A larger cohort of patients with known pathogenic mutations would be necessary to calculate a background/basal rate of falsely flagged variants.

Other groups have attempted to develop comprehensive approaches for variant analysis, analogous to the one proposed here [208–210]. While most employ high-throughput sequencing and classify variants, either the sequences analyzed or the types of variants assessed tend to be limited. In particular, non-coding sequences have not been sequenced or studied to the same extent, and none of these analytical approaches have adopted a common framework for mutation analysis.

Our published oligonucleotide design method [77] produced an average sequence coverage of 98.8%. The capture reagent did not overlap conserved highly repetitive regions, but included divergent repetitive sequences. Nevertheless, neighboring probes

generated reads with partial overlap of repetitive intervals. As previously reported [147], we noted that false positive variant calls within intronic and intergenic regions were the most common consequence of dephasing in low complexity, pyrimidine-enriched intervals. This was not alleviated by processing data with software programs based on different alignment or calling algorithms. Manual review of all intronic or intergenic variants became imperative. As these sequences can still affect functional binding elements detectable by IT analysis (i.e. 3' SSs and SRFBSs), it may prove essential to adopt or develop alignment software that explicitly and correctly identifies variants in these regions [147]. Most variants were confirmed with Sanger sequencing (10/13), and those that could not be confirmed are not necessarily false positives. A recent study demonstrated that NGS can identify variants that Sanger sequencing cannot, and reproducing sequencing results by NGS may be worthwhile before eliminating such variants [211].

4.5 Conclusions

Through a comprehensive protocol based on high-throughput, IT-based and complementary coding sequence analyses, the numbers of VUS can be reduced to a manageable quantity of variants, prioritized by predicted function. While exonic variants corresponded to a small fraction of prioritized variants, there is considerably more evidence for their pathogenicity because clinical sequencing has concentrated in these regions. Our sequencing approach illustrates the importance of sequencing non-coding regions of genes to establish pathogenic mutations not already evident from changes in the amino acid based genetic code [212]. We suggest our approach for variant flagging and prioritization bridges the phase between high-throughput sequencing, variant detection with the time-consuming process of variant classification, including pedigree analysis and functional validation. Subsequent to completion of the present study, ethics approval was obtained for a similar analysis of consented patients with clinical information. This work will be described elsewhere [212].

4.6 References

1. Collins FS, Hamburg MA: First FDA Authorization for Next-Generation Sequencer. N Engl J Med 2013, 369:2369–2371.

2. Green ED, Guyer MS, National Human Genome Research Institute: Charting a course for genomic medicine from base pairs to bedside. Nature 2011, 470:204–213.

3. Cassa CA, Savage SK, Taylor PL, Green RC, McGuire AL, Mandl KD: Disclosing pathogenic genetic variants to research participants: Quantifying an emerging ethical responsibility. Genome Res 2012, 22:421–428.

4. Domchek SM, Bradbury A, Garber JE, Offit K, Robson ME: Multiplex Genetic Testing for Cancer Susceptibility: Out on the High Wire Without a Net? J Clin Oncol 2013, 31:1267–1270.

5. Yorczyk A, Robinson LS, Ross TS: Use of panel tests in place of single gene tests in the cancer genetics clinic. Clin Genet 2015, 88:278–282.

6. Foley SB, Rios JJ, Mgbemena VE, Robinson LS, Hampel HL, Toland AE, Durham L, Ross TS: Use of Whole Genome Sequencing for Diagnosis and Discovery in the Cancer Genetics Clinic. EBioMedicine 2015, 2:74–81.

7. Schwartz GF, Hughes KS, Lynch HT, Fabian CJ, Fentiman IS, Robson ME, Domchek SM, Hartmann LC, Holland R, Winchester DJ, Consensus Conference Committee The International Consensus Conference Committee: Proceedings of the international consensus conference on breast cancer risk, genetics, & risk management, April, 2007. Cancer 2008, 113:2627–2637.

8. Kavanagh D, Anderson HE: Interpretation of genetic variants of uncertain significance in atypical hemolytic uremic syndrome. Kidney Int 2012, 81:11–13.

9. Tavtigian SV, Greenblatt MS, Lesueur F, Byrnes GB, Group IUGVW: In silico analysis of missense substitutions using sequence-alignment based methods. Hum Mutat 2008, 29:1327–1336.

10. Stratton MR, Rahman N: The emerging landscape of breast cancer susceptibility. Nat Genet 2008, 40:17–22.

11. Ready K, Gutierrez-Barrera AM, Amos C, Meric-Bernstam F, Lu K, Hortobagyi G, Arun B: Cancer risk management decisions of women with BRCA1 or BRCA2 variants of uncertain significance. Breast J 2011, 17:210–212.

12. Eggington JM, Bowles KR, Moyes K, Manley S, Esterling L, Sizemore S, Rosenthal E, Theisen A, Saam J, Arnell C, Pruss D, Bennett J, Burbidge LA, Roa B, Wenstrup RJ: A comprehensive laboratory-based program for classification of variants of uncertain significance in hereditary cancer genes. Clin Genet 2014, 86:229–237.

13. Nanda R, Schumm LP, Cummings S, Fackenthal JD, Sveen L, Ademuyiwa F, Cobleigh M, Esserman L, Lindor NM, Neuhausen SL, Olopade OI: Genetic testing in an ethnically diverse cohort of high-risk women: a comparative analysis of BRCA1 and BRCA2 mutations in American families of European and African ancestry. JAMA 2005, 294:1925–1933.

14. U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2012 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute. 2015.

15. Cancer Genome Atlas Network: Comprehensive molecular portraits of human breast tumours. Nature 2012, 490:61–70.

16. Domchek S, Weber BL: Genetic variants of uncertain significance: flies in the ointment. J Clin Oncol Off J Am Soc Clin Oncol 2008, 26:16–17.

17. Braun TA, Mullins RF, Wagner AH, Andorf JL, Johnston RM, Bakall BB, Deluca AP, Fishman GA, Lam BL, Weleber RG, Cideciyan AV, Jacobson SG, Sheffield VC, Tucker BA, Stone EM: Non-exomic and synonymous variants in ABCA4 are an important cause of Stargardt disease. Hum Mol Genet 2013, 22:5136–5145.

18. Castello A, Fischer B, Hentze MW, Preiss T: RNA-binding proteins in Mendelian disease. Trends Genet TIG 2013, 29:318–327.

19. Chatterjee S, Berwal SK, Pal JK: Pathological Mutations in 5′ Untranslated Regions of Human Genes. In eLS. John Wiley & Sons, Ltd; 2001.

20. Halvorsen M, Martin JS, Broadaway S, Laederach A: Disease-associated mutations that alter the RNA structural ensemble. PLoS Genet 2010, 6:e1001074.

21. Misquitta CM, Iyer VR, Werstiuk ES, Grover AK: The role of 3'-untranslated region (3'-UTR) mediated mRNA stability in cardiovascular pathophysiology. Mol Cell Biochem 2001, 224:53–67.

22. Latchman DS: Transcription-Factor Mutations and Disease. N Engl J Med 1996, 334:28–33.

23. Ward AJ, Cooper TA: The Pathobiology of Splicing. J Pathol 2010, 220:152–163.

24. Araujo PR, Yoon K, Ko D, Smith AD, Qiao M, Suresh U, Burns SC, Penalva LOF: Before It Gets Started: Regulating Translation at the 5' UTR. Comp Funct Genomics 2012, 2012:475731.

25. Cáceres JF, Kornblihtt AR: Alternative splicing: multiple control mechanisms and involvement in human disease. Trends Genet TIG 2002, 18:186–193.

26. Teraoka SN, Telatar M, Becker-Catania S, Liang T, Onengüt S, Tolun A, Chessa L, Sanal O, Bernatowska E, Gatti RA, Concannon P: Splicing defects in the ataxia-telangiectasia gene, ATM: underlying mutations and consequences. Am J Hum Genet 1999, 64:1617–1631.

27. Ars E, Serra E, García J, Kruyer H, Gaona A, Lázaro C, Estivill X: Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. Hum Mol Genet 2000, 9:237–247.

28. Paul DS, Soranzo N, Beck S: Functional interpretation of non-coding sequence variation: Concepts and challenges. BioEssays 2014, 36:191–199.

29. Guo Y, Jamison DC: The distribution of SNPs in human gene regulatory regions. BMC Genomics 2005, 6:140.

30. Horvath A, Pakala SB, Mudvari P, Reddy SDN, Ohshiro K, Casimiro S, Pires R, Fuqua SAW, Toi M, Costa L, Nair SS, Sukumar S, Kumar R: Novel insights into breast cancer genetic variance through RNA sequencing. Sci Rep 2013, 3:2256.

31. Pavithra L, Rampalli S, Sinha S, Sreenath K, Pestell RG, Chattopadhyay S: Stabilization of SMAR1 mRNA by PGA2 involves a stem loop structure in the 5' UTR. Nucleic Acids Res 2007, 35:6004–6016.

32. Pérez-Cabornero L, Infante M, Velasco E, Lastra E, Miner C, Durán M: Evaluating the effect of unclassified variants identified in MMR genes using phenotypic features, bioinformatics prediction, and RNA assays. J Mol Diagn JMD 2013, 15:380–390.

33. Zeng T, Dong Z-F, Liu S-J, Wan R-P, Tang L-J, Liu T, Zhao Q-H, Shi Y-W, Yi Y-H, Liao W-P, Long Y-S: A novel variant in the 3' UTR of human SCN1A gene from a patient with Dravet syndrome decreases mRNA stability mediated by GAPDH's binding. Hum Genet 2014, 133:801–811.

34. Gaildrat P, Krieger S, Théry J-C, Killian A, Rousselin A, Berthet P, Frébourg T, Hardouin A, Martins A, Tosi M: The BRCA1 c.5434C->G (p.Pro1812Ala) variant induces a deleterious exon 23 skipping by affecting exonic splicing regulatory elements. J Med Genet 2010, 47:398–403.

35. Tournier I, Vezain M, Martins A, Charbonnier F, Baert-Desurmont S, Olschwang S, Wang Q, Buisine MP, Soret J, Tazi J, Frébourg T, Tosi M: A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. Hum Mutat 2008, 29:1412–1424.

36. Caminsky NG, Mucaki EJ, Rogan PK: Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis. F1000Research 2015, 3:282.

37. Peterlongo P, Catucci I, Colombo M, Caleca L, Mucaki E, Bogliolo M, Marin M, Damiola F, Bernard L, Pensotti V, Volorio S, Dall'Olio V, Meindl A, Bartram C, Sutter C, Surowy H, Sornin V, Dondon M-G, Eon-Marchais S, Stoppa-Lyonnet D, Andrieu N, Sinilnikova OM, Genesis, Mitchell G, James PA, Thompson E, kConFab, Swe-Brca, Marchetti M, Verzeroli C, et al.: FANCM c.5791C>T nonsense mutation (rs144567652) induces exon skipping, affects DNA repair activity and is a familial breast cancer risk factor. Hum Mol Genet 2015, 24:5345–5355.

38. Mucaki EJ, Shirley BC, Rogan PK: Prediction of Mutant mRNA Splice Isoforms by Information Theory-Based Exon Definition. Hum Mutat 2013, 34:557–565.

39. Olsen RKJ, Brøner S, Sabaratnam R, Doktor TK, Andersen HS, Bruun GH, Gahrn B, Stenbroen V, Olpin SE, Dobbie A, Gregersen N, Andresen BS: The ETFDH c.158A>G variation disrupts the balanced interplay of ESE- and ESS-binding proteins thereby causing missplicing and multiple Acyl-CoA dehydrogenation deficiency. Hum Mutat 2014, 35:86–95.

40. Schneider TD, Stormo GD, Yarus MA, Gold L: Delila system tools. Nucleic Acids Res 1984, 12(1 Pt 1):129–140.

41. Schneider TD, Stephens RM: Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 1990, 18:6097–6100.

42. Rogan PK, Faux BM, Schneider TD: Information analysis of human splice site mutations. Hum Mutat 1998, 12:153–171.

43. Chen J-M, Férec C, Cooper DN: A systematic analysis of disease-associated variants in the 3′ regulatory regions of human protein-coding genes II: the importance of mRNA secondary structure in assessing the functionality of 3′ UTR variants. Hum Genet 2006, 120:301–333.

44. Steen K-A, Siegfried NA, Weeks KM: Selective 2’-hydroxyl acylation analyzed by protection from exoribonuclease (RNase-detected SHAPE) for direct analysis of covalent adducts and of nucleotide flexibility in RNA. Nat Protoc 2011, 6:1683–1694.

45. Ferlay J, Shin H-R, Bray F, Forman D, Mathers C, Parkin DM: Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer 2010, 127:2893–2917.

46. Susswein LR, Marshall ML, Nusbaum R, Vogel Postula KJ, Weissman SM, Yackowski L, Vaccari EM, Bissonnette J, Booker JK, Cremona ML, Gibellini F, Murphy PD, Pineda-Alvarez DE, Pollevick GD, Xu Z, Richard G, Bale S, Klein RT, Hruska KS, Chung WK: Pathogenic and likely pathogenic variant prevalence among the first 10,000 patients referred for next-generation cancer panel testing. Genet Med 2015.

47. Levy-Lahad E, Plon SE: Cancer. A risky business--assessing breast cancer risk. Science 2003, 302:574–575.

48. Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, Hogervorst FBL, Hoogerbrugge N, Spurdle AB, Tavtigian SV, IARC Unclassified Genetic Variants Working Group: Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum Mutat 2008, 29:1282–1291.

49. Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, Bishop DT, Weber B, Lenoir G, Chang-Claude J, Sobol H, Teare MD, Struewing J, Arason A, Scherneck S, Peto J, Rebbeck TR, Tonin P, Neuhausen S, Barkardottir R, Eyfjord J, Lynch H, Ponder BA, Gayther SA, Zelada-Hedman M: Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. Am J Hum Genet 1998, 62:676–689.

50. Shah PD, Garber JE, Stopfer JE, Powers J, Nathanson KL, Domchek S: Sensitivity of clinical BRCA1 testing compared with linkage analysis. Volume 30. J Clin Oncol; 2012.

51. Bakker JL, Thirthagiri E, van Mil SE, Adank MA, Ikeda H, Verheul HMW, Meijers-Heijboer H, de Winter JP, Sharan SK, Waisfisz Q: A novel splice site mutation in the noncoding region of BRCA2: implications for Fanconi anemia and familial breast cancer diagnostics. Hum Mutat 2014, 35:442–446.

52. Menéndez M, Castellsagué J, Mirete M, Pros E, Feliubadaló L, Osorio A, Calaf M, Tornero E, Valle J del, Fernández-Rodríguez J, Quiles F, Salinas M, Velasco A, Teulé A, Brunet J, Blanco I, Capellá G, Lázaro C: Assessing the RNA effect of 26 DNA variants in the BRCA1 and BRCA2 genes. Breast Cancer Res Treat 2011, 132:979–992.

53. Borg A, Haile RW, Malone KE, Capanu M, Diep A, Torngren T, Teraoka S, Begg CB, Thomas DC, Concannon P, Mellemkjaer L, Bernstein L, Tellhed L, Xue S, Olson ER, Liang X, Dolle J, Borresen-Dale AL, Bernstein JL: Characterization of BRCA1 and BRCA2 deleterious mutations and variants of unknown clinical significance in unilateral and bilateral breast cancer: the WECARE study. Hum Mutat 2010, 31:E1200–40.

54. Adank MA, Jonker MA, Kluijt I, Mil SE van, Oldenburg RA, Mooi WJ, Hogervorst FBL, Ouweland AMW van den, Gille JJP, Schmidt MK, Vaart AW van der, Meijers-Heijboer H, Waisfisz Q: CHEK2*1100delC homozygosity is associated with a high breast cancer risk in women. J Med Genet 2011, 48:860–863.

55. Baloch AH, Daud S, Raheem N, Luqman M, Ahmad A, Rehman A, Shuja J, Rasheed S, Ali A, Kakar N, Naseeb HK, Mengal MA, Awan MA, Wasim M, Baloch DM, Ahmad J: Missense mutations (p.H371Y, p.D438Y) in gene CHEK2 are associated with breast cancer risk in women of Balochistan origin. Mol Biol Rep 2014, 41:1103–1107.

56. Benusiglio PR, Malka D, Rouleau E, De Pauw A, Buecher B, Noguès C, Fourme E, Colas C, Coulet F, Warcoin M, Grandjouan S, Sezeur A, Laurent-Puig P, Molière D, Tlemsani C, Di Maria M, Byrde V, Delaloge S, Blayau M, Caron O: CDH1 germline mutations and the hereditary diffuse gastric and lobular breast cancer syndrome: a multicentre study. J Med Genet 2013, 50:486–489.

57. Brooks-Wilson AR, Kaurah P, Suriano G, Leach S, Senz J, Grehan N, Butterfield YSN, Jeyes J, Schinas J, Bacani J, Kelsey M, Ferreira P, MacGillivray B, MacLeod P,

Micek M, Ford J, Foulkes W, Australie K, Greenberg C, LaPointe M, Gilpin C, Nikkel S, Gilchrist D, Hughes R, Jackson CE, Monaghan KG, Oliveira MJ, Seruca R, Gallinger S, Caldas C, et al.: Germline E-cadherin mutations in hereditary diffuse gastric cancer: assessment of 42 new families and review of genetic screening criteria. J Med Genet 2004, 41:508–517.

58. Casadei S, Norquist BM, Walsh T, Stray S, Mandell JB, Lee MK, Stamatoyannopoulos JA, King M-C: Contribution of inherited mutations in the BRCA2-interacting protein PALB2 to familial breast cancer. Cancer Res 2011, 71:2222–2229.

59. CHEK2 Breast Cancer Case-Control Consortium: CHEK2*1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. Am J Hum Genet 2004, 74:1175–1182.

60. Garber JE, Offit K: Hereditary cancer predisposition syndromes. J Clin Oncol Off J Am Soc Clin Oncol 2005, 23:276–292.

61. Kangelaris KN, Gruber SB: Clinical implications of founder and recurrent CDH1 mutations in hereditary diffuse gastric cancer. JAMA 2007, 297:2410–2411.

62. Kaurah P, MacMillan A, Boyd N, Senz J, De Luca A, Chun N, Suriano G, Zaor S, Van Manen L, Gilpin C, Nikkel S, Connolly-Wilson M, Weissman S, Rubinstein WS, Sebold C, Greenstein R, Stroop J, Yim D, Panzini B, McKinnon W, Greenblatt M, Wirtzfeld D, Fontaine D, Coit D, Yoon S, Chung D, Lauwers G, Pizzuti A, Vaccaro C, Redal MA, et al.: Founder and recurrent CDH1 mutations in families with hereditary diffuse gastric cancer. JAMA 2007, 297:2360–2372.

63. Kluijt I, Sijmons RH, Hoogerbrugge N, Plukker JT, de Jong D, van Krieken JH, van Hillegersberg R, Ligtenberg M, Bleiker E, Cats A, Dutch Working Group on Hereditary Gastric Cancer: Familial gastric cancer: guidelines for diagnosis, treatment and periodic surveillance. Fam Cancer 2012, 11:363–369.

64. Martin A-M, Kanetsky PA, Amirimani B, Colligon TA, Athanasiadis G, Shih HA, Gerrero MR, Calzone K, Rebbeck TR, Weber BL: Germline TP53 mutations in breast

cancer families with multiple primary cancers: is TP53 a modifier of BRCA1? J Med Genet 2003, 40:e34–e34.

65. Masciari S, Larsson N, Senz J, Boyd N, Kaurah P, Kandel MJ, Harris LN, Pinheiro HC, Troussard A, Miron P, Tung N, Oliveira C, Collins L, Schnitt S, Garber JE, Huntsman D: Germline E-cadherin mutations in familial lobular breast cancer. J Med Genet 2007, 44:726–731.

66. Maxwell KN, Wubbenhorst B, D'Andrea K, Garman B, Long JM, Powers J, Rathbun K, Stopfer JE, Zhu J, Bradbury AR, Simon MS, DeMichele A, Domchek SM, Nathanson KL: Prevalence of mutations in a panel of breast cancer susceptibility genes in BRCA1/2-negative patients with early-onset breast cancer. Genet Med Off J Am Coll Med Genet 2015, 17:630–638.

67. Minion LE, Dolinsky JS, Chase DM, Dunlop CL, Chao EC, Monk BJ: Hereditary predisposition to ovarian cancer, looking beyond BRCA1/BRCA2. Gynecol Oncol 2015, 137:86–92.

68. Olivier M, Goldgar DE, Sodha N, Ohgaki H, Kleihues P, Hainaut P, Eeles RA: Li-Fraumeni and related syndromes: correlation between tumor type, family structure, and TP53 genotype. Cancer Res 2003, 63:6643–6650.

69. Pharoah PD, Guilford P, Caldas C, International Gastric Cancer Linkage Consortium: Incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families. Gastroenterology 2001, 121:1348–1353.

70. Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, Reid S, Spanova K, Barfoot R, Chagtai T, Jayatilake H, McGuffog L, Hanks S, Evans DG, Eccles D, Breast Cancer Susceptibility Collaboration (UK), Easton DF, Stratton MR: PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nat Genet 2007, 39:165–167.

71. Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, North B, Jayatilake H, Barfoot R, Spanova K, McGuffog L, Evans DG, Eccles D, Breast Cancer

Susceptibility Collaboration (UK), Easton DF, Stratton MR, Rahman N: ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nat Genet 2006, 38:873–875.

72. Sidransky D, Tokino T, Helzlsouer K, Zehnbauer B, Rausch G, Shelton B, Prestigiacomo L, Vogelstein B, Davidson N: Inherited p53 gene mutations in breast cancer. Cancer Res 1992, 52:2984–2986.

73. Slater EP, Langer P, Niemczyk E, Strauch K, Butler J, Habbe N, Neoptolemos JP, Greenhalf W, Bartsch DK: PALB2 mutations in European familial pancreatic cancer families. Clin Genet 2010, 78:490–494.

74. Thompson D, Duedal S, Kirner J, McGuffog L, Last J, Reiman A, Byrd P, Taylor M, Easton DF: Cancer risks and mortality in heterozygous ATM mutation carriers. J Natl Cancer Inst 2005, 97:813–822.

75. Tischkowitz M, Capanu M, Sabbaghian N, Li L, Liang X, Vallée MP, Tavtigian SV, Concannon P, Foulkes WD, Bernstein L, WECARE Study Collaborative Group, Bernstein JL, Begg CB: Rare germline mutations in PALB2 and breast cancer risk: a population-based study. Hum Mutat 2012, 33:674–680.

76. Walsh T, Casadei S, Coats KH, Swisher E, Stray SM, Higgins J, Roach KC, Mandell J, Lee MK, Ciernikova S, Foretova L, Soucek P, King M-C: Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. JAMA 2006, 295:1379–1388.

77. Dorman SN, Shirley BC, Knoll JHM, Rogan PK: Expanding probe repertoire and improving reproducibility in human genomic hybridization. Nucleic Acids Res 2013, 41:e81.

78. Pinkel D, Landegent J, Collins C, Fuscoe J, Segraves R, Lucas J, Gray J: Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. Proc Natl Acad Sci U S A 1988, 85:9138–9142.

79. Smit A, Hubley R, Green P: RepeatMasker Open-4.0. 2013-2015.

80. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C: Solution Hybrid Selection with Ultra-long Oligonucleotides for Massively Parallel Targeted Sequencing. Nat Biotechnol 2009, 27:182–189.

81. Chou H-H, Hsia A-P, Mooney DL, Schnable PS: Picky: oligo microarray design for large genomes. Bioinforma Oxf Engl 2004, 20:2893–2902.

82. Markham NR, Zuker M: UNAFold: software for nucleic acid folding and hybridization. Methods Mol Biol Clifton NJ 2008, 453:3–31.

83. Zuker M: Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 2003, 31:3406–3415.

84. Predictive Cancer Genetics Steering Committee Ontario physicians' guide to referral of patients with family history of cancer to a familial cancer genetics clinic or genetics clinic. Ont Med Rev 2001, 68:24–30.

85. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ: A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011, 43:491–498.

86. Philippe N, Salson M, Commes T, Rivals E: CRAC: an integrated approach to the analysis of RNA-seq reads. Genome Biol 2013, 14:R30.

87. Picard [http://picard.sourceforge.net/]. Accessed June 1, 2015.

88. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010, 20:1297–1303.

89. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: Integrative genomics viewer. Nat Biotechnol 2011, 29:24–26.

90. Thorvaldsdóttir H, Robinson JT, Mesirov JP: Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013, 14:178–192.

91. Shirley BC, Mucaki EJ, Whitehead T, Costea PI, Akan P, Rogan PK: Interpretation, stratification and evidence for sequence variants affecting mRNA splicing in complete human genome sequences. Genomics Proteomics Bioinformatics 2013, 11:77–85.

92. Mutation Forecaster [https://www.mutationforecaster.com/index.php]. Accessed June 1, 2015.

93. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A: Information content of binding sites on nucleotide sequences. J Mol Biol 1986, 188:415–431.

94. Dhir A, Buratti E: Alternative splicing: role of pseudoexons in human disease and potential therapeutic strategies. FEBS J 2010, 277:841–855.

95. Tavanez JP, Madl T, Kooshapur H, Sattler M, Valcárcel J: hnRNP A1 proofreads 3’ splice site recognition by U2AF. Mol Cell 2012, 45:314–329.

96. Paradis C, Cloutier P, Shkreta L, Toutant J, Klarskov K, Chabot B: hnRNP I/PTB can antagonize the splicing repressor activity of SRp30c. RNA N Y N 2007, 13:1287–1300.

97. ENCODE Project Consortium: An integrated encyclopedia of DNA elements in the human genome. Nature 2012, 489:57–74.

98. Boggs K, Reisman D: Increased p53 transcription prior to DNA synthesis is regulated through a novel regulatory element within the p53 promoter. Oncogene 2005, 25:555–565.

99. Chen Y, Xu J, Borowicz S, Collins C, Huo D, Olopade OI: c-Myc activates BRCA1 gene expression through distal promoter elements in breast cancer cells. BMC Cancer 2011, 11:246.

100. Gueven N, Keating K, Fukao T, Loeffler H, Kondo N, Rodemann HP, Lavin MF: Site-directed mutagenesis of the ATM promoter: Consequences for response to proliferation and ionizing radiation. Genes Chromosomes Cancer 2003, 38:157–167.

101. Frietze S, Wang R, Yao L, Tak YG, Ye Z, Gaddis M, Witt H, Farnham PJ, Jin VX: Cell type-specific binding patterns reveal that TCF7L2 can be tethered to the genome by association with GATA3. Genome Biol 2012, 13:R52.

102. Connor AE, Baumgartner RN, Baumgartner KB, Kerber RA, Pinkston C, John EM, Torres-Mejia G, Hines L, Giuliano A, Wolff RK, Slattery ML: Associations between TCF7L2 polymorphisms and risk of breast cancer among Hispanic and non-Hispanic white women: the Breast Cancer Health Disparities Study. Breast Cancer Res Treat 2012, 136:593–602.

103. Burwinkel B, Shanmugam KS, Hemminki K, Meindl A, Schmutzler RK, Sutter C, Wappenschmidt B, Kiechle M, Bartram CR, Frank B: Transcription factor 7-like 2 (TCF7L2) variant is associated with familial breast cancer risk: a case-control study. BMC Cancer 2006, 6:268.

104. Chen J, Yuan T, Liu M, Chen P: Association between TCF7L2 Gene Polymorphism and Cancer Risk: A Meta-Analysis. PLoS ONE 2013, 8:e71730.

105. Purrington KS, Slager S, Eccles D, Yannoukakos D, Fasching PA, Miron P, Carpenter J, Chang-Claude J, Martin NG, Montgomery GW, Kristensen V, Anton-Culver H, Goodfellow P, Tapper WJ, Rafiq S, Gerty SM, Durcan L, Konstantopoulou I, Fostira F, Vratimos A, Apostolou P, Konstanta I, Kotoula V, Lakis S, Dimopoulos MA, Skarlos D, Pectasides D, Fountzilas G, Beckmann MW, Hein A, et al.: Genome-wide association study identifies 25 known breast cancer susceptibility loci as risk factors for triple-negative breast cancer. Carcinogenesis 2014, 35:1012–1019.

106. Bi C, Rogan PK: Bipartite pattern discovery by entropy minimization-based multiple local alignment. Nucleic Acids Res 2004, 32:4979–4991.

107. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, Rando OJ, Birney E, Myers RM, Noble WS, Snyder M, Weng Z: Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res 2012, 22:1798–1812.

108. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D: GeneCards: integrating information about genes, proteins and diseases. Trends Genet TIG 1997, 13:163.

109. Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR: RBPDB: a database of RNA-binding specificities. Nucleic Acids Res 2011, 39(Database issue):D301–8.

110. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, Na H, Irimia M, Matzat LH, Dale RK, Smith SA, Yarosh CA, Kelly SM, Nabet B, Mecenas D, Li W, Laishram RS, Qiao M, Lipshitz HD, Piano F, Corbett AH, Carstens RP, Frey BJ, Anderson RA, Lynch KW, Penalva LOF, et al.: A compendium of RNA-binding motifs for decoding gene regulation. Nature 2013, 499:172–177.

111. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng H, Goity A, van Bakel H, Lozano J-C, Galli M, Lewsey MG, Huang E, Mukherjee T, Chen X, Reece-Hoyes JS, Govindarajan S, Shaulsky G, Walhout AJM, Bouget F-Y, Ratsch G, Larrondo LF, Ecker JR, Hughes TR: Determination and inference of eukaryotic transcription factor sequence specificity. Cell 2014, 158:1431–1443.

112. Dayem Ullah AZ, Lemoine NR, Chelala C: SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). Nucleic Acids Res 2012, 40:W65–W70.

113. Dayem Ullah AZ, Lemoine NR, Chelala C: A practical guide for the functional annotation of genetic variations using SNPnexus. Brief Bioinform 2013, 14:437–447.

114. Chelala C, Khan A, Lemoine NR: SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. Bioinformatics 2009, 25:655–661.

115. dbSNP [http://www.ncbi.nlm.nih.gov/SNP/]. Accessed June 1, 2015.

116. Exome Variant Server [http://evs.gs.washington.edu/EVS/]. Accessed June 1, 2015.

117. 1000Genomes [http://www.1000genomes.org/]. Accessed June 1, 2015.

118. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: A method and server for predicting damaging missense mutations. Nat Methods 2010, 7:248–249.

119. Reva B, Antipin Y, Sander C: Determinants of protein function revealed by combinatorial entropy optimization. Genome Biol 2007, 8:R232.

120. Reva B, Antipin Y, Sander C: Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 2011, 39:e118.

121. Choi Y: A Fast Computation of Pairwise Sequence Alignment Scores Between a Protein and a Set of Single-locus Variants of Another Protein. In Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine. New York, NY, USA: ACM; 2012:414–417. [BCB ’12]

122. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP: Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS ONE 2012, 7:e46688.

123. ClinVar [http://www.ncbi.nlm.nih.gov/clinvar/]. Accessed June 1, 2015.

124. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR: ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 2013:gkt1113.

125. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NST, Abeysinghe S, Krawczak M, Cooper DN: Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 2003, 21:577–581.

126. Human Gene Mutation Database (HGMD) [http://hgmd/cf/ac/uk/ac/index.php]. Accessed June 1, 2015.

127. Fokkema IFAC, Taschner PEM, Schaafsma GCP, Celli J, Laros JFJ, den Dunnen JT: LOVD v.2.0: the next generation in gene variant databases. Hum Mutat 2011, 32:557–563.

128. Leiden Open Variation Database (LOVD) - Ataxia Telangiectasia Mutated (ATM) [http://chromium.lovd.nl/LOVD2/variants.php?action=search_unique&select_db=ATM]. Accessed June 1, 2015.

129. LOVD - IARC Breast Cancer Type 1 susceptibility protein (BRCA1) [http://brca.iarc.fr/LOVD/variants.php?action=view_unique&select_db=BRCA1]. Accessed June 1, 2015.

130. LOVD - IARC Breast Cancer Type 2 susceptibility protein (BRCA2) [http://brca.iarc.fr/LOVD/variants.php?action=view_unique&select_db=BRCA2]. Accessed June 1, 2015.

131. LOVD - Leiden Open Variation Database Partner and localizer of BRCA2 (FANCN) (PALB2) [https://grenada.lumc.nl/LOVD2/shared1/variants.php?action=search_unique&select_db=PALB2]. Accessed June 1, 2015.

132. LOVD - Leiden Open Variation Database tumour protein p53 (TP53) [http://proteomics.bio21.unimelb.edu.au/lovd/variants/TP53]. Accessed June 1, 2015.

133. Zhejiang University Center for Genetic and Genomic Medicine (ZJU-CGGM) cadherin 1, type 1, E-cadherin (epithelial) (CDH1) [http://www.genomed.org/lovd2/variants.php?action=search_unique&select_db=CDH1]. Accessed June 1, 2015.

134. Zhejiang University Center for Genetic and Genomic Medicine (ZJU-CGGM) checkpoint kinase 2 (CHEK2) [http://www.genomed.org/lovd2/variants.php?action=search_unique&select_db=CHEK2]. Accessed June 1, 2015.

135. Domain Mapping of Disease Mutations (DM2) [http://bioinf.umbc.edu/dmdm]. Accessed June 1, 2015.

136. Expert Protein Analysis System (ExPASy) [http://www.expasy.org/]. Accessed June 1, 2015.

137. The UniProt Consortium: UniProt: a hub for protein information. Nucleic Acids Res 2015, 43:D204–D212.

138. UniProt [http://uniprot.org/]. Accessed June 1, 2015.

139. Breast Cancer Information Core (BIC) Database [https://research.nhgri.nih.gov/projects/bic/Member/index/shtml]. Accessed June 1, 2015.

140. Evidence-based Network for the Interpretation of Germline Mutant Alleles (ENIGMA) [http://enigmaconsortium.org/]. Accessed June 1, 2015.

141. International Agency for Research on Cancer (IARC) TP53 Database [http://p53.iarc.fr/tp53genevariations.aspx]. Accessed June 1, 2015.

142. Ozcelik H, Knight JA, Glendon G, Yazici H, Carson N, Ainsworth PJ, Taylor S a. M, Feilotter H, Carter RF, Boyd NF, Andrulis IL, Ontario Cancer Genetics Network: Individual and family characteristics associated with protein truncating BRCA1 and BRCA2 mutations in an Ontario population based series from the Cooperative Family Registry for Breast Cancer Studies. J Med Genet 2003, 40:e91.

143. Maier C, Herkommer K, Luedeke M, Rinckleb A, Schrader M, Vogel W: Subgroups of familial and aggressive prostate cancer with considerable frequencies of BRCA2 mutations. The Prostate 2014, 74:1444–1451.

144. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER: BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 2009, 6:677–681.

145. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Ch'ang L-Y, Huang W, Liu B, Shen Y, Tam PK-H, Tsui L-C, Waye MMY, Wong JT-F, Zeng C, Zhang Q, Chee MS, Galver LM, Kruglyak S, Murray SS, Oliphant AR, Montpetit A, Hudson TJ, Chagnon F, Ferretti V, Leboeuf M, Phillips MS, Verner A, Kwok P-Y, Duan S, et al.: The International HapMap Project. Nature 2003, 426:789–796.

146. McIver LJ, Fondon III JW, Skinner MA, Garner HR: Evaluation of microsatellite variation in the 1000 Genomes Project pilot studies is indicative of the quality and utility of the raw data and alignments. Genomics 2011, 97:193–199.

147. Tae H, Kim D-Y, McCormick J, Settlage RE, Garner HR: Discretized Gaussian mixture for genotyping of microsatellite loci containing homopolymer runs. Bioinformatics 2014, 30:652–659.

148. Castéra L, Krieger S, Rousselin A, Legros A, Baumann J-J, Bruet O, Brault B, Fouillet R, Goardon N, Letac O, Baert-Desurmont S, Tinat J, Bera O, Dugast C, Berthet P, Polycarpe F, Layet V, Hardouin A, Frébourg T, Vaur D: Next-generation sequencing for the diagnosis of hereditary breast and ovarian cancer using genomic capture targeting multiple candidate genes. Eur J Hum Genet EJHG 2014, 22:1305–1313.

149. Austen B, Barone G, Reiman A, Byrd PJ, Baker C, Starczynski J, Nobbs MC, Murphy RP, Enright H, Chaila E, Quinn J, Stankovic T, Pratt G, Taylor AMR: Pathogenic ATM mutations occur rarely in a subset of multiple myeloma patients. Br J Haematol 2008, 142:925–933.

150. Ding H, Mao C, Li S-M, Liu Q, Lin L, Chen Q: Lack of association between ATM C.1066-6T > G mutation and breast cancer risk: a meta-analysis of 8,831 cases and 4,957 controls. Breast Cancer Res Treat 2011, 125:473–477.

151. Chen J, Guo K, Kastan MB: Interactions of nucleolin and ribosomal protein L26 (RPL26) in translational control of human p53 mRNA. J Biol Chem 2012, 287:16467–16476.

152. Champion-Arnaud P, Reed R: The prespliceosome components SAP 49 and SAP 145 interact in a complex implicated in tethering U2 snRNP to the branch site. Genes Dev 1994, 8:1974–1983.

153. Li YI, Sanchez-Pulido L, Haerty W, Ponting CP: RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. Genome Res 2015, 25:1–13.

154. Dobričić J, Krivokuća A, Brotto K, Mališić E, Radulović S, Branković-Magić M: Serbian high-risk families: extensive results on BRCA mutation spectra and frequency. J Hum Genet 2013, 58:501–507.

155. Nelson AC, Holt JT: Impact of RING and BRCT domain mutations on BRCA1 protein stability, localization and recruitment to DNA damage. Radiat Res 2010, 174:1–13.

156. Clark SL, Rodriguez AM, Snyder RR, Hankins GDV, Boehning D: Structure-Function Of The Tumor Suppressor BRCA1. Comput Struct Biotechnol J 2012, 1.

157. Leung CCY, Glover JNM: BRCT domains: easy as one, two, three. Cell Cycle Georget Tex 2011, 10:2461–2470.

158. Håkansson S, Johannsson O, Johansson U, Sellberg G, Loman N, Gerdes AM, Holmberg E, Dahl N, Pandis N, Kristoffersson U, Olsson H, Borg A: Moderate frequency of BRCA1 and BRCA2 germ-line mutations in Scandinavian familial breast cancer. Am J Hum Genet 1997, 60:1068–1078.

159. Scottish/Northern Irish BRCAI/BRCA2 Consortium: BRCA1 and BRCA2 mutations in Scotland and Northern Ireland. Br J Cancer 2003, 88:1256–1262.

160. Hussain S, Wilson JB, Medhurst AL, Hejna J, Witt E, Ananth S, Davies A, Masson J-Y, Moses R, West SC, de Winter JP, Ashworth A, Jones NJ, Mathew CG: Direct interaction of FANCD2 with BRCA2 in DNA damage response pathways. Hum Mol Genet 2004, 13:1241–1248.

161. Chang YF, Imam JS, Wilkinson MF: The nonsense-mediated decay RNA surveillance pathway. Annu Rev Biochem 2007, 76:51–74.

162. Oliveira C, Senz J, Kaurah P, Pinheiro H, Sanges R, Haegert A, Corso G, Schouten J, Fitzgerald R, Vogelsang H, Keller G, Dwerryhouse S, Grimmer D, Chin S-F, Yang H-K, Jackson CE, Seruca R, Roviello F, Stupka E, Caldas C, Huntsman D: Germline CDH1 deletions in hereditary diffuse gastric cancer families. Hum Mol Genet 2009, 18:1545–1555.

163. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J: A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014, 46:310–315.

164. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, et al.: Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008, 456:53–59.

165. Platinum Genomes [http://www.illumina.com/platinumgenomes/]. Accessed July 31, 2015.

166. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL, ACMG Laboratory Quality Assurance Committee: Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med Off J Am Coll Med Genet 2015, 17:405–424.

167. Tavtigian SV, Greenblatt MS, Goldgar DE, Boffetta P, IARC Unclassified Genetic Variants Working Group: Assessing pathogenicity: overview of results from the IARC Unclassified Genetic Variants Working Group. Hum Mutat 2008, 29:1261–1264.

168. Easton DF, Deffenbaugh AM, Pruss D, Frye C, Wenstrup RJ, Allen-Brady K, Tavtigian SV, Monteiro ANA, Iversen ES, Couch FJ, Goldgar DE: A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. Am J Hum Genet 2007, 81:873–883.

169. Viner C, Dorman SN, Shirley BC, Rogan PK: Validation of predicted mRNA splicing mutations using high-throughput transcriptome data. F1000Research 2014, 3:8.

170. Carleton KL: Quantification of transcript levels with quantitative RT-PCR. Methods Mol Biol Clifton NJ 2011, 772:279–295.

171. ENIGMA BRCA1/2 Gene Variant Classification Criteria, v1.1 [http://enigmaconsortium.org/documents/ENIGMA_Rules_2015-03-26.pdf]. Accessed June 1, 2015.

172. Deveci M, Catalyürek UV, Toland AE: mrSNP: software to detect SNP effects on microRNA binding. BMC Bioinformatics 2014, 15:73.

173. Vlachos IS, Paraskevopoulou MD, Karagkouni D, Georgakilas G, Vergoulis T, Kanellos I, Anastasopoulos I-L, Maniou S, Karathanou K, Kalfakakou D, Fevgas A, Dalamagas T, Hatzigeorgiou AG: DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. Nucleic Acids Res 2015, 43(Database issue):D153–159.

174. Mucaki EJ, Ainsworth P, Rogan PK: Comprehensive prediction of mRNA splicing effects of BRCA1 and BRCA2 variants. Hum Mutat 2011, 32:735–742.

175. Bisio A, Nasti S, Jordan JJ, Gargiulo S, Pastorino L, Provenzani A, Quattrone A, Queirolo P, Bianchi-Scarrà G, Ghiorzo P, Inga A: Functional analysis of CDKN2A/p16INK4a 5'-UTR variants predisposing to melanoma. Hum Mol Genet 2010, 19:1479–1491.

176. Berry JA, Cervantes-Sandoval I, Nicholas EP, Davis RL: Dopamine is required for learning and forgetting in Drosophila. Neuron 2012, 74:530–542.

177. Sribudiani Y, Metzger M, Osinga J, Rey A, Burns AJ, Thapar N, Hofstra RMW: Variants in RET associated with Hirschsprung's disease affect binding of transcription factors and gene expression. Gastroenterology 2011, 140:572–582.e2.

178. Knebelmann B, Forestier L, Drouot L, Quinones S, Chuet C, Benessy F, Saus J, Antignac C: Splice-mediated insertion of an Alu sequence in the COL4A3 mRNA causing autosomal recessive Alport syndrome. Hum Mol Genet 1995, 4:675–679.

179. Wallace MR, Andersen LB, Saulino AM, Gregory PE, Glover TW, Collins FS: A de novo Alu insertion results in neurofibromatosis type 1. Nature 1991, 353:864–866.

180. Wiestner A, Tehrani M, Chiorazzi M, Wright G, Gibellini F, Nakayama K, Liu H, Rosenwald A, Muller-Hermelink HK, Ott G, Chan WC, Greiner TC, Weisenburger DD, Vose J, Armitage JO, Gascoyne RD, Connors JM, Campo E, Montserrat E, Bosch F, Smeland EB, Kvaloy S, Holte H, Delabie J, Fisher RI, Grogan TM, Miller TP, Wilson WH, Jaffe ES, Staudt LM: Point mutations and genomic deletions in CCND1 create stable truncated cyclin D1 mRNAs that are associated with increased proliferation rate and shorter survival. Blood 2007, 109:4599–4606.

181. Lévesque É, Bélanger A-S, Harvey M, Couture F, Jonker D, Innocenti F, Cecchin E, Toffoli G, Guillemette C: Refining the UGT1A Haplotype Associated with Irinotecan-Induced Hematological Toxicity in Metastatic Colorectal Cancer Patients Treated with 5-Fluorouracil/Irinotecan-Based Regimens. J Pharmacol Exp Ther 2013, 345:95–101.

182. Fujiwara Y, Minami H: An overview of the recent progress in irinotecan pharmacogenetics. Pharmacogenomics 2010, 11:391–406.

183. Palomaki GE, Bradley LA, Douglas MP, Kolor K, Dotson WD: Can UGT1A1 genotyping reduce morbidity and mortality in patients with metastatic colorectal cancer treated with irinotecan? An evidence-based review. Genet Med 2009, 11:21–34.

184. Biesecker LG: Opportunities and challenges for the integration of massively parallel genomic sequencing into clinical practice: Lessons from the ClinSeq^{TM} project. Genet Med Off J Am Coll Med Genet 2012, 14:393–398.

185. Schneider TD: Information content of individual genetic sequences. J Theor Biol 1997, 189:427–441.

186. Bonnet-Dupeyron M-N, Combes P, Santander P, Cailloux F, Boespflug-Tanguy O, Vaurs-Barrière C: PLP1 splicing abnormalities identified in Pelizaeus-Merzbacher disease and SPG2 fibroblasts are associated with different types of mutations. Hum Mutat 2008, 29:1028–1036.

187. Fei J: Splice Site Mutation-Induced Alteration of Selective Regional Activity Correlates with the Role of a Gene in Cardiomyopathy. J Clin Exp Cardiol 2013, S12:004.

188. Khan SG, Metin A, Gozukara E, Inui H, Shahlavi T, Muniz-Medina V, Baker CC, Ueda T, Aiken JR, Schneider TD, Kraemer KH: Two essential splice lariat branchpoint sequences in one intron in a xeroderma pigmentosum DNA repair gene: mutations result in reduced XPC mRNA levels that correlate with cancer risk. Hum Mol Genet 2004, 13:343–352.

189. von Kodolitsch Y, Berger J, Rogan PK: Predicting severity of haemophilia A and B splicing mutations by information analysis. Haemoph Off J World Fed Hemoph 2006, 12:258–262.

190. Martoni E, Urciuolo A, Sabatelli P, Fabris M, Bovolenta M, Neri M, Grumati P, D'Amico A, Pane M, Mercuri E, Bertini E, Merlini L, Bonaldo P, Ferlini A, Gualandi F: Identification and characterization of novel collagen VI non-canonical splicing mutations causing Ullrich congenital muscular dystrophy. Hum Mutat 2009, 30:E662–672.

191. Nasim MT, Ogo T, Ahmed M, Randall R, Chowdhury HM, Snape KM, Bradshaw TY, Southgate L, Lee GJ, Jackson I, Lord GM, Gibbs JSR, Wilkins MR, Ohta-Ogo K, Nakamura K, Girerd B, Coulet F, Soubrier F, Humbert M, Morrell NW, Trembath RC,

Machado RD: Molecular genetic characterization of SMAD signaling molecules in pulmonary arterial hypertension. Hum Mutat 2011, 32:1385–1389.

192. Pink AE, Simpson MA, Desai N, Dafou D, Hills A, Mortimer P, Smith CH, Trembath RC, Barker JNW: Mutations in the γ-secretase genes NCSTN, PSENEN, and PSEN1 underlie rare forms of hidradenitis suppurativa (acne inversa). J Invest Dermatol 2012, 132:2459–2461.

193. Sanders DA, Ross-Innes CS, Beraldi D, Carroll JS, Balasubramanian S: Genome-wide mapping of FOXM1 binding reveals co-binding with estrogen receptor alpha in breast cancer cells. Genome Biol 2013, 14:R6.

194. Suga Y, Tsuda T, Nagai M, Sakaguchi Y, Jitsukawa O, Yamamoto M, Hitomi K, Yamanishi K: Lamellar ichthyosis with pseudoexon activation in the transglutaminase 1 gene. J Dermatol 2015, 42:642–645.

195. Rogan PK, Svojanovsky S, Leeder JS: Information theory-based analysis of CYP2C19, CYP2D6 and CYP3A5 splicing mutations. Pharmacogenetics 2003, 13:207–218.

196. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler and D: The Human Genome Browser at UCSC. Genome Res 2002, 12:996–1006.

197. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: GenBank: update. Nucleic Acids Res 2004, 32(Database issue):D23–26.

198. Rogan P, Mucaki E: Population Fitness and Genetic Load of Single Nucleotide Polymorphisms Affecting mRNA splicing. ArXiv11070716 Q-Bio 2011.

199. Di L-J, Fernandez AG, De Siervi A, Longo DL, Gardner K: Transcriptional regulation of BRCA1 expression by a metabolic switch. Nat Struct Mol Biol 2010, 17:1406–1413.

200. Wu X, Webster SR, Chen J: Characterization of tumor-associated Chk2 mutations. J Biol Chem 2001, 276:2971–2974.

201. Durocher D, Henckel J, Fersht AR, Jackson SP: The FHA domain is a modular phosphopeptide recognition motif. Mol Cell 1999, 4:387–394.

202. Bell DW, Varley JM, Szydlo TE, Kang DH, Wahrer DC, Shannon KE, Lubratovich M, Verselis SJ, Isselbacher KJ, Fraumeni JF, Birch JM, Li FP, Garber JE, Haber DA: Heterozygous germ line hCHK2 mutations in Li-Fraumeni syndrome. Science 1999, 286:2528–2531.

203. Varley JM, Evans DG, Birch JM: Li-Fraumeni syndrome--a molecular and clinical review. Br J Cancer 1997, 76:1–14.

204. Lee SB, Kim SH, Bell DW, Wahrer DC, Schiripo TA, Jorczak MM, Sgroi DC, Garber JE, Li FP, Nichols KE, Varley JM, Godwin AK, Shannon KM, Harlow E, Haber DA: Destabilization of CHK2 by a missense mutation associated with Li-Fraumeni Syndrome. Cancer Res 2001, 61:8062–8067.

205. Biswas DK, Shi Q, Baily S, Strickland I, Ghosh S, Pardee AB, Iglehart JD: NF-kappa B activation in human breast cancer specimens and its role in cell proliferation and apoptosis. Proc Natl Acad Sci U S A 2004, 101:10137–10142.

206. Antoniou AC, Easton DF: Models of genetic susceptibility to breast cancer. Oncogene 2006, 25:5898–5905.

207. Peto J: Breast cancer susceptibility—A new look at an old model. Cancer Cell 2002, 1:411–412.

208. Kurian AW, Hare EE, Mills MA, Kingham KE, McPherson L, Whittemore AS, McGuire V, Ladabaum U, Kobayashi Y, Lincoln SE, Cargill M, Ford JM: Clinical Evaluation of a Multiple-Gene Sequencing Panel for Hereditary Cancer Risk Assessment. J Clin Oncol 2014, 32:2001–2009.

209. Kassahn KS, Scott HS, Caramins MC: Integrating Massively Parallel Sequencing into Diagnostic Workflows and Managing the Annotation and Clinical Interpretation Challenge. Hum Mutat 2014, 35:413–423.

210. Li M-X, Gui H-S, Kwan JSH, Bao S-Y, Sham PC: A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Res 2012, 40:e53.

211. Kluska A, Balabas A, Paziewska A, Kulecka M, Nowakowska D, Mikula M, Ostrowski J: New recurrent BRCA1/2 mutations in Polish patients with familial breast/ovarian cancer detected by next generation sequencing. BMC Med Genomics 2015, 8:19.

212. Caminsky NG, Eliseos JM, Perri AM., Lu R, Knoll JHM, Rogan PK: Prioritizing variants in complete Hereditary Breast and Ovarian Cancer genes in patients lacking known *BRCA* mutations, Hum Mutat 2016, 37:640–652.

Chapter 5

5 Prioritizing Variants in Complete Hereditary Breast and Ovarian Cancer Genes in Patients Lacking Known *BRCA* Mutations

The work presented in this chapter is reproduced (with permission, Appendix D.1) from:

Caminsky,N.G., Eliseos,J.M., Perri,A.M., Lu,R., Knoll,J.H.M., Rogan,P.K. (2016) Prioritizing variants in complete Hereditary Breast and Ovarian Cancer genes in patients lacking known *BRCA* mutations. *Hum. Mutat.*, 37, 640–652

5.1 Introduction

Currently, the lifetime risk for a woman to develop breast cancer (BC) is 12.3% and 1.3% in the case of ovarian cancer (OC [Howlander et al., 2014]). Approximately 5-10% of all BC cases are hereditary in nature, versus 25% for OC, where relative risk (RR) of BC or OC with one affected 1st degree family member is estimated at 2.1 and 3.1, respectively [Stratton et al., 1998; Walsh et al., 2011]. Two highly penetrant genes, *BRCA1* and *BRCA2*, are associated with a large proportion of HBOC cases. However, the estimated rate of linkage to these genes is significantly higher than the proportion of pathogenic mutations identified in HBOC families [Ford et al., 1998], suggesting unrecognized or unidentified variants in *BRCA1/2*.

Clinical *BRCA1/2* testing is restricted primarily to coding regions. Limitations on how variants can be interpreted, lack of functional validation, and mutations in other genes contribute to uninformative results. The heritability that is not associated with *BRCA* genes is likely due to other genetic factors rather than environmental causes, specifically moderate- and low-risk susceptibility genes [Antoniou and Easton, 2006]. Hollestelle et al. (2010) point out the challenges in estimating increased risks associated with mutations in these genes, as the disease patterns are often incompletely penetrant, and require large pedigree studies to confidently assess pathogenicity.

Next-generation sequencing (NGS) of gene panels for large cohorts of affected and unaffected individuals has become an increasingly popular approach to confront these challenges. Numerous HBOC gene variants have been catalogued, including cases in which RR has been determined; however the literature is also flooded with variants lacking a clinical interpretation [Cassa et al., 2012]. It is not feasible to functionally evaluate the effects all of the VUS identified by NGS and *in silico* assessment of variants is often limited to structural changes or concerns evolutionary conservation among species. Several approaches have been developed to better assess variants from exome and genome-wide NGS data [Duzkale et al., 2013; Kircher et al., 2014]. Nevertheless, there is an unmet need for other methods that quickly and accurately bridge variant identification and classification.

To begin to address this problem, we sought to provide potentially novel interpretations of noncoding sequence changes, based on disruption or acquisition of interactions with proteins that recognize nucleic acid binding sites. Information theory (IT) based analysis predicts changes in sequence binding affinity, and it has been applied and validated for use in the analysis of splice sites (SSs), SRBSs [Rogan et al., 1998, 2003; Mucaki et al., 2013; Caminsky et al., 2015] and TFBSs [Gadiraju et al., 2003]. A unified framework based on IT requires binding genome-scale site data devoid of consensus sequence bias [Schneider, 1997], for example, photoactivatable-ribonucleotide-enhanced cross-linking and immunoprecipitation (PAR-CLIP), ChIP-Seq, and a comprehensive, validated set of SSs. Although these data sources are heterogeneous, the IT models and binding site affinities derived from them are uniformly scaled (in units of bits). Thus, binding interactions involving disparate proteins or other recognition molecules can be measured and directly compared.

We have described a unified IT framework for the identification and prioritization of variants in coding and noncoding region s of *BRCA1*, *BRCA2*, and five other HBOC genes (*ATM*, *CDH1*, *CHEK2*, *PALB2*, and *TP53* [Mucaki et al., 2016]. This approach was applied to a cohort of 102 individuals lacking BRCA mutations with a history of HBOC. This distinguished prioritized variants from flagged alleles conferring small

changes to regulatory protein binding site sequences in 70.6% of cases [Mucaki et al., 2016].

In the present study, we have sequenced 13 additional genes that have been deemed HBOC susceptibility loci (*BARD1*, *EPCAM*, *MLH1*, *MRE11A*, *MSH2*, *MSH6*, *MUTYH*, *NBN*, *PMS2*, *PTEN*, *RAD51B*, *STK11*, and *XRCC2* [Minion et al., 2015]). These genes encode proteins with roles in DNA repair, surveillance, and cell cycle regulation (Figure 5.1; for further evidence supporting this gene set see Appendix D.2 [Apostolou and Fostira 2013; Al Bakir and Gabra 2014]), and are associated with specific disease syndromes that confer an increased risk of BC and OC, as well as many other types of cancer (Appendix D.3). High-risk genes confer > 4-times increased risk of BC compared to the general population. *BRCA1* and *BRCA2* are estimated to increase risk 20-fold [Antoniou et al., 2003]. Pathogenic variants in other high-risk genes, *CDH1*, *PTEN*, *STK11*, and *TP53*, are rarely seen outside of their associated syndromes, and account for < 1% of hereditary BC cases [Maxwell and Domchek, 2013]. *EPCAM*, *MLH1*, *MSH2*, *MSH6*, and *PMS2* have also been proposed to harbor high-risk BC alleles, but the RR is still controversial [Maxwell and Domchek, 2013]. Genes with moderate-risk alleles, *ATM*, *CHEK2*, and *PALB2*, cause between a 2- and 4-fold increased risk of BC [Apostolou and Fostira, 2013; Maxwell and Domchek, 2013]. The remaining genes (*BARD1*, *MRE11A*, *MUTYH*, *NBN*, *RAD51B*, and *XRCC2*) are newly identified and currently associated with unknown risks for HBOC (Figure 5.1).

We report NGS of hybridization-enriched, complete genic and surrounding regions of these genes, followed by variant analysis in 287 consented patients from Southwestern Ontario, Canada with previously uninformative HBOC test results. (Except for 6 individuals, these patients were different from our submitted study). We then reduced the set of potentially pathogenic gene variants in each individual by prioritizing the results of coding and IT-analyses. After applying a frequency-based filter, the IT-based framework prioritizes variants based on their predicted effect on the recognition of sequence elements involved in mRNA splicing, transcription, and untranslated region (UTR) binding, combined with UTR secondary structure and coding variant analysis. Our approach integrates disparate sources of information, including bioinformatic analyses,

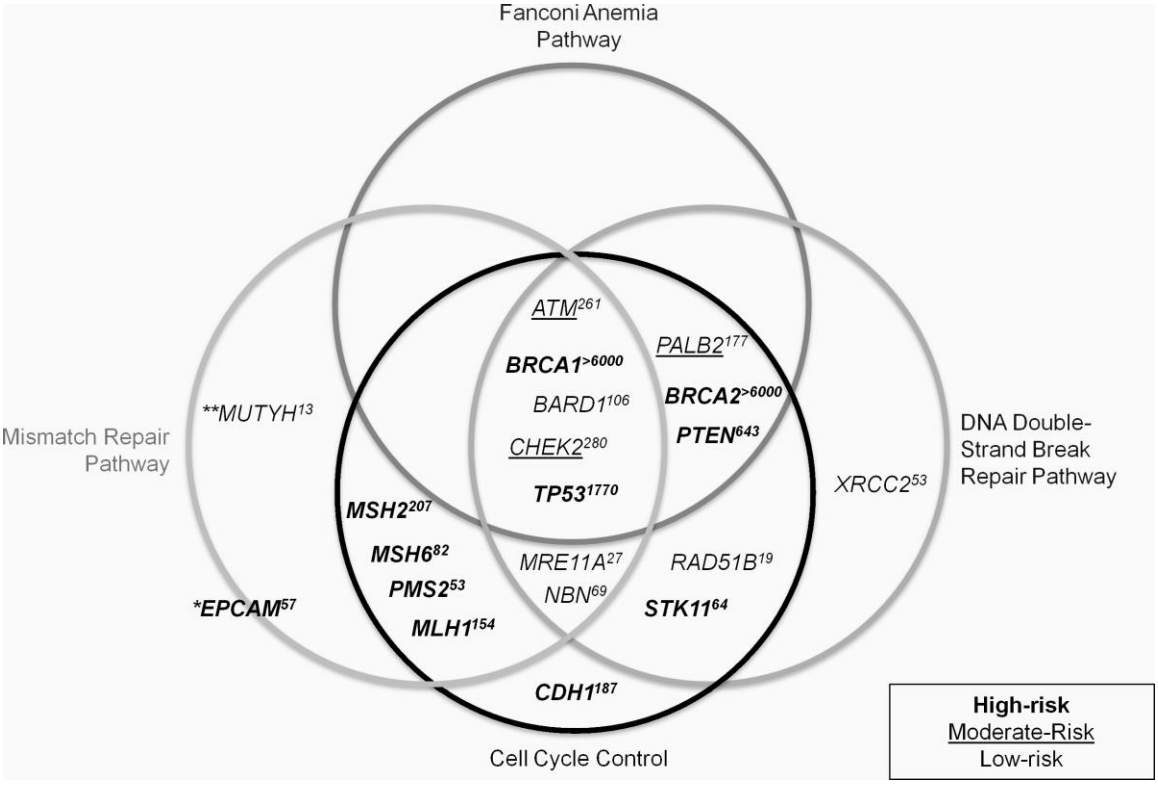

Figure 5.1: Common genomic pathways among 20 HBOC genes, including risk and relevant literature. The left, top, and right circles indicate sequenced genes that play important roles in the mismatch repair (MMR), Fanconi anemia, and DNA double-strand break repair pathways, respectively. The bottom circle contains genes involved in cell cycle control. Genes considered to present a high risk of breast and/or ovarian cancer when mutated are bolded, moderate-risk genes are underlined, and low-risk genes are in normal font. The estimated number of articles listing a gene's association with breast or ovarian cancer (based on a systematic search in PubMed [performed June 2015]) is indicated in superscript. ∗∗ *MUTYH* is only high risk in the case of bi-allelic mutations. ∗ *EPCAM* is not involved in any pathways, but is associated with hereditary nonpolyposis colorectal cancer (HNPCC) by virtue of the fact that 3' deletions of *EPCAM* can cause epigenetic silencing of *MSH2*, causing Lynch syndrome protein. See Appendix D.1 for citations and further evidence supporting this gene set.

likelihood ratios based on familial segregation, allele frequencies, and published findings to prioritize disease-associated mutation candidates.

5.2 Methods

5.2.1 Ethics and Patient Recruitment

Recruitment and consent of human participants was approved by the University of Western Ontario Research Ethics Board (Protocol 103746). Patients were enrolled from January, 2014 through March, 2015 at London Health Sciences Centre (LHSC). Patients met the following criteria: male or female, aged between 25 and 75 years, > 10% risk of having an inherited mutation in a breast/ovarian cancer gene, diagnosed with BC and/or OC, and previously receiving uninformative results for a known, pathogenic *BRCA1* or *BRCA2* variant in either the patient or other relatives (by Protein Truncation Test [PTT] and/or Multiplex Ligation-dependent Probe Amplification [MLPA]).

The median age of onset for patients (N = 287; Appendix D.5-Supp. Fig. S1) with BC was 48 (N = 277), and 46 for OC (N = 17), and 7 were diagnosed with both BC and OC. Furthermore, 31 patients had bilateral BC (98 patients at diagnosis; 23 developed tumors on the opposite side after the initial occurrence), 1 had bilateral OC, and 13 have had recurrent BC in the same breast. There was a single case of male BC (Appendix D.4).

5.2.2 Probe Design, Sample Preparation, and Sequencing

Probes for sequence capture were designed by *ab initio* single copy analysis, as described in Mucaki et al. [2016] and Dorman et al. [2013]. The probes covered1,103,029 nt across the 21 sequenced genes, including the negative control gene *ATP8B1* (see Appendix D.5 for gene names, GenBank accession numbers, and OMIM reference numbers). This set of genes was proposed for evaluation at the evidence-based network for the Interpretation of Germline Mutant Alleles (ENIGMA) Consortium Meeting (2013). Other genes that have been found to be mutated in HBOC could not be included (e.g. *BRIP1*, *RAD50*, *RAD51C*, *RAD51D* [Heikkinen et al. 2003; Seal et al. 2006; Janatova et al. 2015]).

Patient DNA extracted from peripheral blood was either obtained from the initial genetic testing at LHSC Molecular Genetics Laboratory or isolated from recent samples. NGS libraries were prepared using modifications to a published protocol (Gnirke et al., 2009) described in Mucaki et al. [2016], and all post-capture pull-down steps were automated

(Appendix D.5). An Illumina Genome Analyzer IIx instrument in our laboratory was used for sequencing.

Library preparation and re-sequencing were repeated for samples with initial average coverage below our minimum threshold (< 30x). To ensure that the proper sample was re-sequenced, the variant call format (VCF) files from each run were compared to all others in the run using VCF-compare (http://vcftools.sourceforge.net/). VCF files from separate runs for the re-sequenced patient were concordant, except for minor differences in variant call rates due to differences in coverage. The aligned reads from both runs were then merged (with BAMtools; http://sourceforge.net/projects/bamtools/).

Samples were demultiplexed and aligned using CASAVA (Consensus Assessment of Sequencing and Variation; v1.8.2 [DePristo et al. 2011]) and CRAC (Complex Reads Analysis & Classification; v1.3.0; http://crac.gforge.inria.fr/). Aligned BAM files were then pre-processed for variant calling with Picard [v.1.109; http://broadinstitute.github.io/picard/] (MarkDuplicates, AddorReplaceReadGroups, FixMateInformation). The Genome Analysis Toolkit (GATK v3.1; http://www.broadinstitute.org/gatk/) was then used for variant calling using the modules 'Indel realigner' and the 'Unified Genotyper'. Variants flagged by bioinformatic analysis [see *Variant Analysis* below] were also assessed by manual inspection with the Integrative Genome Viewer v2.3 (IGV; http://www.broadinstitute.org/igv/). Variants in this study are written in HGVS notation, are based on cDNA sequence, and comply with journal guidelines.

5.2.3 Information Models

Models for natural splice sites (SSs) and splicing regulatory factors (SRFs) are described in Mucaki et al. [2013]. These models were used to predict deleterious effects on natural splicing, the activation of cryptic SSs, and changes to binding of splicing enhancers and silencers. In addition, using a combination of cryptic site activation and hnRNPA1 site prediction, pseudoexon formation was also assessed.

We previously built models for TFBSs (N = 83) using ENCODE ChIP-seq data [ENCODE Project Consortium, 2012; Mucaki et al., 2016]. Due to the inclusion of the additional genes, eight additional transcription factors (TFs) were identified from the literature and ENCODE ChIP-seq data from BC cell lines with evidence of binding and potentially regulating these genes. However, models for three of these TFs passed our quality control criteria (TFIIIB150 [*BDP1*], PBX3 and ZNF274; described in Lu et al. [2017]. Appendix D.6 contains the full list of TFs (N = 86) and indicates which genes exhibit evidence of promoter or other binding events. Noise models (N = 5), reflecting motifs of interacting cofactors or sequence-specific histone modifying events, were excluded (Appendix D.5).

Information weight matrices, $R_i(b, l)$, for sequences bound by RNA-binding proteins (RBPs) were derived from frequency matrices published in the Catalog of Inferred Sequence Binding Preferences of RNA binding protein (CISBP-RNA; http://cisbp-rna.ccbr.utoronto.ca/) and RNA-Binding Protein Database (BPDB; http://rbpdb.ccbr.utoronto.ca/). These $R_i(b, l)$s were used to compute changes in binding affinity due to SNVs, using conservative minimum information thresholds described in Mucaki et al. [2016]. Finally, predicted changes in UTR structure resulting from variants were determined using SNPfold [http://ribosnitch.bio.unc.edu/snpfold/;Halvorsenetal.,2010]. Significant changes in UTR structure and stability were represented using mfold (http://unafold.rna.albany.edu/?q = mfold).

5.2.4 Variant Analysis

Information analysis has been used in the interpretation of variant effects on binding sites containing these changes, whether this involves the creation or strengthening, or the abolition or weakening of a site [Rogan et al., 1998]. This analysis was applied to all variants identified by NGS. Changes in information are directly related to changes in thermodynamic entropy and thus binding affinity [Rogan et al., 1998]. For example, a 1.0 bit change in information corresponds to at least a twofold change in binding affinity. Information theoretical analysis of SSs and SRF binding sites has been extensively used

and proven to be reliable and robust (85.2% accuracy when compared to variants validated by expression studies) [Caminsky et al., 2015].

Information analysis was automated and thresholds for changes were applied programmatically based on our previously validated criteria [Rogan et al., 1998, 2003; von Kodolitschetal., 2006; Dorman et al., 2014]. This reduced manual review of prioritized variants, databases, and the literature. A minimum 1.0 bit threshold was set for variants predicted to affect natural SSs or that activate a cryptic SS by exceeding the strength of cognate natural sites. Variants affecting splicing regulatory, transcription, and RBP binding sites were assessed more stringently and had a minimum threshold of 4.0 bits, i.e. $\geq$ 16-fold, in order to be flagged for further assessment. A population frequency filter was also applied to variants with allele frequencies >1% (in the NCBI Short Genetic Variations database (dbSNP)) or >5% of our patient cohort. Such variants were eliminated from further consideration.

To assess coding changes affecting predicted protein chain length or amino acid(s) composition, we used SNPnexus (http://hsnpnexus.org/). Insertion/deletions (indels) and nonsense mutations were noted, and missense variants were further assessed with in silico tools (Mutation Assessor, http://mutationassessor.org/; PolyPhen2, http://genetics.bwh.arvard.edu/pph2/; PROVEAN/SIFT, http://provean.jcvi.org/) by referencing the published literature and consulting mutation databases (listed in Appendix D.7; see Mucaki et al. [2016] for more details on variant analysis). Variants remained prioritized unless there was clear evidence (co-segregation analysis or functional assays) supporting the nonpathogenicity of the variant.

EPCAM mutations in familial cancer are limited to 3' deletions causing epigenetic silencing of *MSH2*, and there is currently no evidence of other types of variants that alter its mRNA transcript or protein product [Ligtenberg et al., 2009]. Therefore, with the exception of indels, none of the variants flagged in *EPCAM* were prioritized. We chose to prioritize variants in *MUTYH* using the same framework as all other genes, despite *MUTYH* pathogenicity resulting from biallelic variants [Jones et al., 2002], because it is possible that a second *MUTYH* mutation remains unrecognized.

All protein truncating (nonsense and indels) as well as potentially pathogenic splicing and missense mutations were Sanger sequenced for confirmation (details in Appendix D.8).

5.2.5 Negative Control

Variants present in the *ATP8B1* gene were used as negative controls for our variant analysis framework. Initially, it was included in the list of prioritized HBOC genes provided by ENIGMA, but evidence for its association with HBOC is lacking in the published literature. Furthermore, it is not a known susceptibility gene for any type of cancer (mutations in *ATP8B1* cause progressive familial intrahepatic cholestasis [Gonzales et al., 2014]) and is infrequently mutated in breast tumors in several studies (e.g., see Cancer Genome Atlas Network [2012]).

5.2.6 Likelihood Ratios

Patients with prioritized coding and/or splicing variants, which we consider the most likely to be pathogenic, were selected for co-segregation analysis (N = 24) using an online tool that calculates the likelihood of a variant being deleterious based on pedigree information (https://www.msbi.nl/cosegregation/; Mohammadi et al. [2009]). Genotypes were assigned based on phenotype such that family members with breast or OC at any age were assigned the same genotype as the patient in our study (“carrier”) and family members affected by other cancers, other diseases, or who are disease free were assigned the “noncarrier” genotype. Because the penetrance parameters cannot be altered from the settings given for *BRCA1* or *BRCA2*, the *BRCA2* option was selected for patients with prioritized variants in non-*BRCA* genes. Penetrance in *BRCA2* is known to be lower than *BRCA1* values [Mohammadi et al., 2009]. Current evidence suggests that mutations in non-*BRCA* genes may be less penetrant than those in the *BRCA* genes [Apostolou and Fostira, 2013]; however, the penetrance of many of these variants remains unknown (Appendix D.5).

5.3 Results

5.3.1 Variant Analysis

We identified 38,372 unique variants among 287 patients (26,636 intronic, 7,287 intergenic, and 714 coding), on average 1,975 variants per patient, before any filtering criteria were applied. The extensive span of sequences captured in this study, that is, complete genes and flanking regions, constrained the genomic density and sequence coverage that could be achieved; this precluded accurate copy number estimation based solely on read counts.

5.3.1.1 Natural Site Variants

The Shannon Human Splicing Mutation Pipeline (http://www.mutationforecaster.com; Shirley et al. [2013]) was used to predict the effect of the 14,458 variants that could potentially affect splicing, of which 244 reduced natural SS strength. Further stringent filtering of the natural SS based on information content changes and allele frequency resulted in seven flagged variants (Appendix D.9). Henceforth, allele frequency of known variants can be found in their associated supplemental table (where available).

Four of these variants were prioritized (Table 5.1). A novel synonymous variant in exon 2 of *RAD51B*, c.84G>A (p.Gln28 =), is predicted to increase exon skipping by weakening the natural splice donor ($R_{i,final}$ = 5.2 bits, ΔR_i = –3.0 bits). A known *ATM* variant, c.6198+1G>A (8-1D.9-1B [Stankovic et al., 1998]), abolishes the natural donor SS of constitutively spliced exon 42 ($R_{i,final}$ = –13.7 bits, ΔR_i = –18.6 bits). There is no evidence in public databases for appreciable alternative splicing of this exon in normal breast tissues. The variant will either lead to exon skipping or activation of a preexisting cryptic site (Figure 5.2). An ataxia-telangiectasia patient with this variant exhibited low expression, protein truncation, and abolished kinase activity of ATM [Reiman et al., 2011]. *MLH1* c.306+4A>G causes increased exon skipping (and a decrease in wild-type exon relative expression) due to the weakening ($R_{i,final}$ = 6.0 bits, ΔR_i = –2.6 bits) of the exon 3 natural donor. Tournier et al. [2008] assessed this variant using an *ex vivo* splicing assay and observed cryptic site activation and exon 3 skipping. *MRE11A* c.2070+2A>T is indicated in ClinVar as likely pathogenic and abolishes the natural donor site of exon 19

Table 5.1: Prioritized Variants Predicted by IT to Affect Natural and Cryptic Splicing

Gene	Variant	rsID (dbSNP142) Allele Frequency (%)[c]	Information Change			Consequence
			$R_{i,initial}$ (bits)	$R_{i,final}$ (bits)	ΔR_i (bits)	
ATM	NM_000051.3:c.6198+1G>A [Stankovic et al., 1998; Reiman et al., 2011]	-	4.9	-13.7	-18.6	Abolished natural[d,g]
MRE11A	NM_005591.3:c.2070+2A>T[a]	-	7.6	-11	-18.6	Abolished natural[d,g]
MLH1	NM_000249.2:c.306+4A>G[a] [Tournier et al., 2008]	rs267607733	8.6	6	-2.6	Weakened natural[e]
RAD51B	NM_002877.4:c.84G>A[a] p.Gln28=	Novel	8.2	5.2	-3	Weakened natural[d]
BARD1	NM_000465.2:c.1454C>T[a] p.Ala485Val	Novel	-2.7	4.4	7.1	Created cryptic[e]
BRCA1	NM_007294.2:c.5074+107C>T	rs373676607	-1.3	5.7	7	Created cryptic[f,h]
CDH1	NM_004360.3:c.1223C>G[a] p.Ala408Gly [Schrader et al. 2011]	Novel	-0.6	4.3	4.9	Created cryptic[e]

RAD51B	NM_002877.4:c.958-29A>T[b]	rs34436700 0.78	2.2	4.4	2.2	Strengthened cryptic[f]
STK11	NM_000455.4:c.375-194GT>AC	rs35113943 17.61 rs117211142 0.80	7.5	8.8	1.3	Strengthened cryptic[f]
XRCC2	NM_005431.1:c.122-154G>T	Novel	8.1	10	1.9	Strengthened cryptic[f]

[a] Confirmed by Sanger sequencing

[b] Ambiguous Sanger sequencing results

[c] If available

[d] Exon skipping

[e] Exon truncation

[f] Intron retention

[g] Use of alternate isoform

[h] Reduced expression of natural isoform

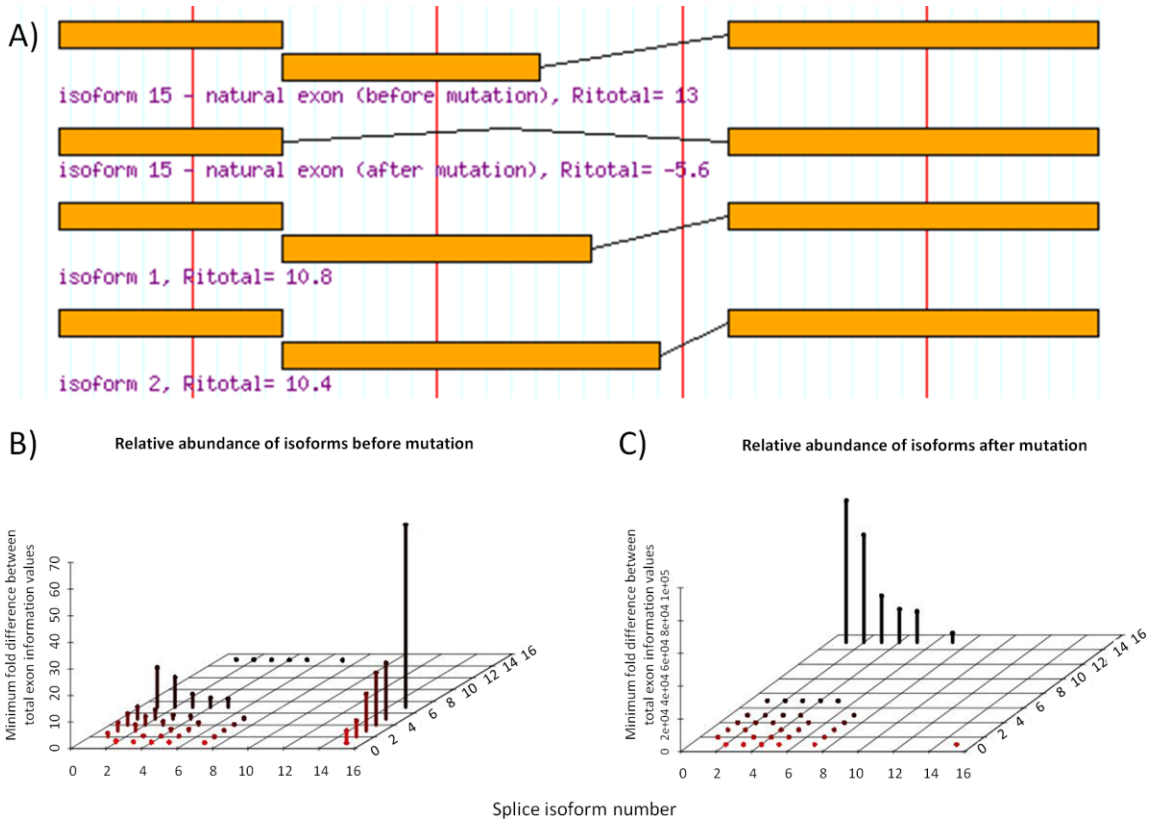

Figure 5.2: Predicted isoforms and relative abundance as a consequence of ATM natural splice variant c.6198+1G>A. (A) Intronic *ATM* variant c.6198+1G>A abolishes the natural donor of exon 42 ($\boldsymbol{R_{i,initial}}$ = 4.9 bits, $\Delta\boldsymbol{R_i}$ = –18.6 bits), and would either result in exon skipping (causing a frame-shift; isoform 15 after mutation), or possibly activate a downstream cryptic site (isoform 1 maintains reading frame, isoform 2 would not). (B) The peaks in plot display the predicted abundance (Y-axis) of a splice isoform (X-axis) relative to another predicted isoform (Z-axis). In the wild-type mRNA, the natural exon (isoform 15) has the highest predicted relative abundance. Before mutation, it is predicted to be approximately fivefold stronger than isoform 1 and 2. (C) After mutation, isoform 1 and 2 is now > 100,000-fold stronger than isoform 15 (abolished wild-type exon). Isoform 2 to be slightly less abundant than 1.

($R_{i,final}$ = –11.0 bits, ΔR_i = –18.6 bits), while strengthening a cryptic site 5 nt upstream of the splice junction ($R_{i,final}$ = 8.1 bits, ΔR_i = 0.6 bits). Either cryptic SS activation or complete exon skipping are predicted.

The *BRCA2* variant c.68-7T>A was not prioritized, as its pathogenicity has not been proven. While there is evidence that this variant induces (in-frame) exon skipping [Théry et al., 2011], it did not segregate with disease in HBOC pedigrees, where abnormal splicing was not seen [Santos et al., 2014]. The *ATM* variant c.1066-6T>G, previously reported in Mucaki et al. [2016], was also not prioritized as the variant does not correlate with BC risk [Ding et al., 2011].

5.3.1.2 Activation of Cryptic Splicing

The Shannon Pipeline identified 9,480 variants that increased the strength of at least one cryptic site, of which nine met or exceeded the defined thresholds for information change. Six of these were prioritized (Table 5.1). A novel *BARD1* variant in exon 6 (c.1454C>T; p.Ala485Val) creates a donor SS ($R_{i,final}$ = 4.4 bits, ΔR_i = 7.1 bits), which would produce a 58 nt frame shifted exon if activated. The natural donor SS of exon 6, 116 nt downstream of the variant, is stronger (5.5 bits), but the Automated Splice Site and Exon Definition Analysis (ASSEDA, http://mutationforecaster.com) server predicts equal levels of expression of both natural and cryptic exons. A *BRCA1* mutation 5074+107C>T downstream of exon 16 is predicted to extend the exon by 105 nt and be slightly more abundant than the natural exon ($R_{i,final}$ of 8.6 and 8.1 bits, respectively). *CDH1* c.1223C>G (p.Ala408Gly), previously reported in a *BRCA*-negative lobular BC patient with no family history of gastric cancer [Schrader et al., 2011], creates a cryptic donor site ($R_{i,final}$ = 4.3 bits, ΔR_i = 4.9 bits) in exon 9, 97 nt downstream of the natural acceptor. While residual splicing of the normal exon is still expected, the cryptic is predicted to become the predominant splice form (~twice as abundant).

STK11 c.375-194GT>AC (rs35113943 and rs117211142) and the novel *XRCC2* c.122-154G>T both strengthen strong preexisting cryptic sites exceeding the $R_{i,total}$ values of their respective natural exons. Finally, a known *RAD51B* variant 29 nt upstream of exon

10: c.958-29A>T strengthens a cryptic acceptor site ($R_{i,final}$ = 4.4 bits, ΔR_i = 2.2 bits) that, if activated, would produce a transcript retaining 21 intronic nucleotides.

The remaining cryptic site variants (Appendix D.9) were not prioritized. The novel *BRCA2* c.7618-269_7618-260del10 variant is predicted to create a cryptic site with an exon having a lower $R_{i,total}$ value (5.2 bits) than the natural exon (6.6 bits). *PMS2* c.1688G>T (p.Arg563Leu; rs63750668; three patients) does not segregate with disease. Drost et al. [2013] demonstrated that this variant does not impair DNA repair activity. Finally, *RAD51B* c.728A>G (p.Lys243Arg; rs34594234; 7 patients) predicts an increase in the abundance of the cryptic exon; however, the natural exon remains the predominant isoform.

5.3.1.3 Pseudoexon Activation

Pseudoexons arise from creation or strengthening of an intronic cryptic SS in close proximity to another intron site of opposite polarity. Our analysis detected 623 variants with such intronic cryptic sites, of which 17 were prioritized (among nine genes), occurring within 250 nt of a preexisting site of opposite polarity, with an hnRNPA1 site within 5 nt of the acceptor of the predicted pseudoexon (Appendix D.10). Three are novel (*BRCA2* c.7007+824C>T, *BRCA2* c.8332-1130G>T, and *PTEN* c.802-796C>A) and the remainder were present in dbSNP. Seven of these variants (*BARD1* c.1315-168C>T, *BRCA2* c.631+271A>G, *MLH1* c.1559-1732A>T, *MRE11A* c.1783+2259A>G, *MSH6* c.260+1758G>A, *PTEN* c.79+4780C>T, and *RAD51B* c.1037-1012C>A), although rare, occur in multiple patients, and one patient had predicted pseudoexons in both *BARD1* and *RAD51B*.

5.3.1.4 SRF Binding

Variants within exons or within 500 nt of a natural SS (*N* = 9,998) were assessed for their potential effect on SRF binding sites (SRFBSs). Initially 216 unique variants were flagged (Appendix D.11), but after considering each in the context of the SRF function and location within the gene [Caminsky et al., 2015], we prioritized 148, of which 57 are novel. Some prioritized variants affect distant SRFs that may activate cryptic sites, but were not predicted to affect natural splicing. Of the 88 suitable prioritized variants for

which exon definition analysis was performed (where initial $R_{i,total}$ of the exon > SRF gap surprisal value), 55 were predicted to induce or contribute to increased exon skipping. For example, an uncommon *ATM* missense variant within exon 41, c.6067G>A (p.Gly2023Arg; rs11212587), strengthens an hnRNPA1 site ($R_{i,final}$ = 5.2 bits, ΔR_i = 4.7 bits) 30 nt from the natural donor, and is predicted to induce exon 41 skipping ($\Delta R_{i,total}$ = –9.5 bits).

5.3.1.5 TF Binding

To assess potential changes to TFBSs, variants occurring from 10 kb upstream of the start of transcription through the end of the first intron were analyzed by IT, flagging 88 (of 4,530 identified; Appendix D.12). Considering the gene context of each TFBS and extent of information change, we prioritized 36 variants. The following example illustrates the rationale for highlighting these variants: *BRCA1* c.-19-433A>G abolishes a binding site for HSF 1 ($R_{i,initial}$ = 5.5 bits, ΔR_i = –7.8 bits). While HSF 1 is known to be a transcriptional activator associated with poor BC prognosis [Santagata et al., 2011], the specific effect of reduced HSF 1 binding to BRCA1 has not been established. Similarly, *MLH1* c.-4285T>C (rs115211110; five patients) significantly weakens a C/EBPβ site ($R_{i,initial}$ = 10.1 bits, ΔR_i = –6.3 bits), a TF that has been shown to play a role in BC development and progression [Zahnow, 2009]. Another *MLH1* variant, c.-6585T>C (novel), greatly decreases the binding strength ($R_{i,initial}$ = 12.5 bits, ΔR_i = –10.8 bits) of the NF-κB p65 subunit, which is activated in ER-negative breast tumors [Biswas et al., 2004]. Two prioritized variants (*PMS2* c.-9059G>C and *XRCC2* c.-163C>A) weaken PAX5 binding sites, a TF which when overexpressed can result in mammary carcinoma cells regaining epithelial cell characteristics [Vidal et al., 2010].

5.3.1.6 Alterations to mRNA Structure

A total of 1,355 variants were identified in the 5’ and 3’ UTRs of the patients. Analysis of these variants with SNPfold flagged three unique variants ($P < 0.05$) in *BRCA1*, *BARD1*, and *XRCC2* (Table 5.2). The predicted mRNA 2° structures of the reference and variant sequences are shown in Figure 5.3 (generated with mfold). The *BRCA1* variant occurs in the 3’ UTR of all known transcript isoforms (NM_007294.3:c.∗1332T>C;

Table 5.2: Variants Predicted by SNPfold to Significantly Affect UTR Structure

Gene	Variant	UTR Position	rsID (dbSNP142) Allele Frequency (%)[a]	Rank	*p-value*
BARD1	XM_005246728.1: c.-53G>T (c.33G>T p.Gln11His)	5'UTR	rs143914387 0.04	6/600	0.01
BRCA1	NM_007294.3:c.*1332T>C NM_007299.3:c.*1438T>C	3'UTR	rs8176320 0.42	13/450	0.03
XRCC2	NM_005431.1:c.-76C>T	5'UTR	rs547538731 0.08	3/300	0.01

[a] If known.

rs8176320; 3 patients). The most likely inferred structure consisting of a short arm and a larger stem loop is destabilized when the variant nucleotide is present (Figure 5.3A and B). The *BARD1* variant falls within the 5' UTR of a rare isoform (XM_005246728.1:c.-53G>T; rs143914387; five patients) and is within the coding region of a more common transcript (NM_000465.2:c.33G>T; p.Gln11His). While the top ranked isoform following mutation is similar to the wild-type structure, the second-ranked isoform (ΔG = +1.88kcal/mol) is distinctly different, creating a loop in a long double-stranded structure (Figure 5.3C and D). The *XRCC2* variant is within its common 5' UTR (NM_005431.1:c.-76C>T) and is located 11 nt downstream from the 5' end of the mRNA. The variant nucleotide disrupts a potential GC base pair, leading to a large stem-loop that could allow access for binding of several RBPs (Figure 5.3E and F). The variant simultaneously strengthens PUM2 ($R_{i,initial}$ = 2.8 bits, ΔR_i = 4.4 bits, positions 11–17) and RBM28 sites ($R_{i,initial}$ = 4.0, ΔR_i = 3.6bits, positions10–13); however, there is a stronger NCL site (8.3 bits, positions 20–31) in the area that is not affected and may compete for binding.

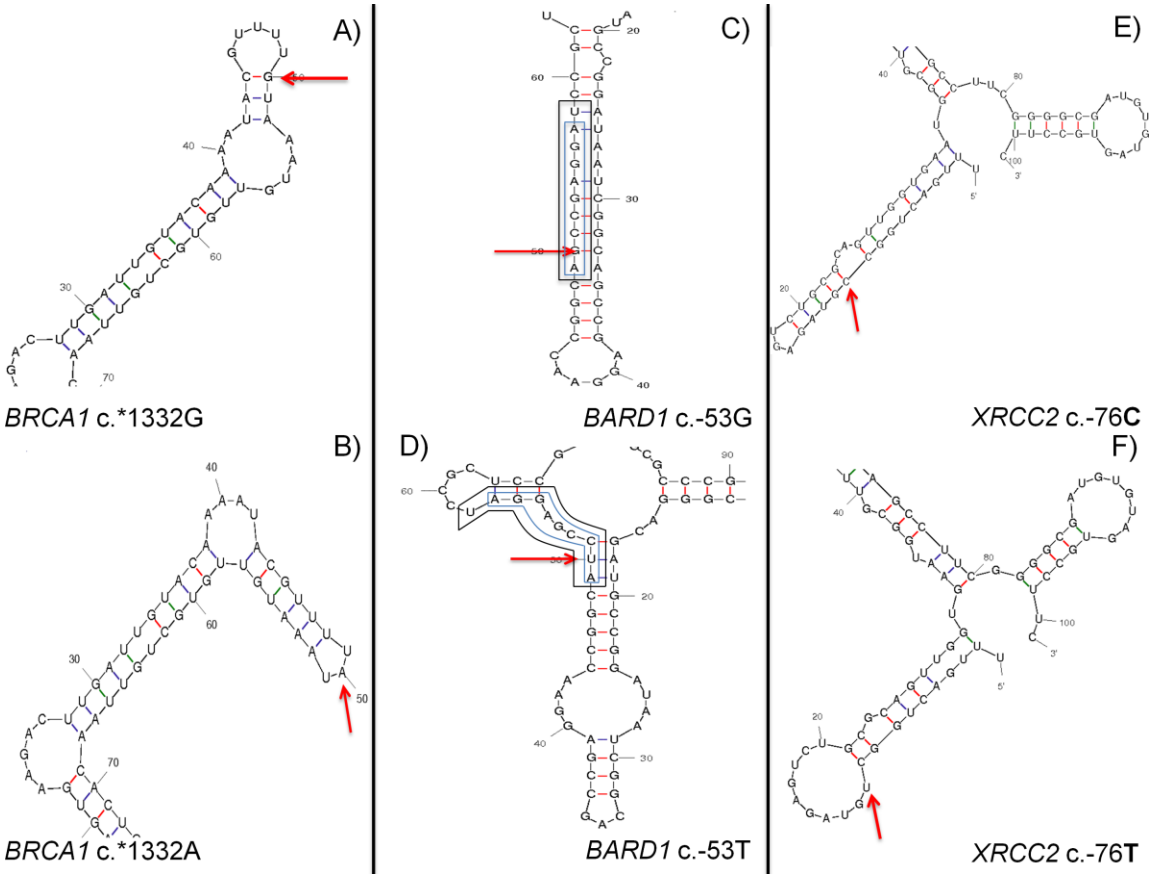

Figure 5.3: Predicted RNA structure change due to variants flagged by SNPfold using mfold. Wild-type (A, C, and E) and variant (B, D, and F) structures are displayed. The variant nucleotide is marked with an arrow. (A) Predicted wild-type structure of *BRCA1* 3' UTR surrounding c.∗1332G>A. (B) *BRCA1* 3' UTR structure due to c.∗1332A variant, extending arm length while reducing hairpin size. (C) BARD1 5' UTR structure of rare isoform (XM_005246728.1:c.-53G>T). Two overlapping preexisting RBP sites (SRSF7 [outer box] and SRSF2 [inner box]) are predicted and either could occupy this location if accessible. (D) 2° *BARD1* 5' UTR structure of the region predicted only with sequence containing the c.-53T mutation. The primary predicted c.-53T structure is identical to wild type (with one disrupted C-G bond leading to a 4.1 kcal/mol lower $\Delta \boldsymbol{G}$). The variant both weakens and abolishes the preexisting SRSF7 and SRSF2 sites, respectively. (E) *XRCC2* structure within common 5' UTR surrounding c.-76C>T variant. (F) *XRCC2* 5' UTR structure predicted from c.-76T sequence, containing a hairpin not found in wild type. This hairpin may allow for the binding of previously inaccessible nucleotides including the altered nucleotide.

5.3.1.7 RBP Binding

Using IT models of 76 RBBSs, 33 UTR variants were prioritized (Appendix D.13) from the initial list of 1,367 UTR variants. Interestingly, one of the three variants that destabilized the mRNA was also flagged using our RBP scan. The *BARD1* c.-53A>C variant weakens a predicted 8.3 bit SRSF7 site (ΔR_i = –3.0 bits) while simultaneously abolishing a predicted 9.7 bit SRSF2 site (ΔR_i = –29.7 bits) (Figure 5.3C and D).

5.3.2 Exonic Protein-Altering Variants

5.3.2.1 Protein Truncating

Of the 714 identified coding variants, six were indels, each of which were found in a single patient, and two preserved the reading frame. Two indels were novel (*BRCA1*:c.3550_3551insA [p.Gly1184Glufs] and *CDH1*:c.30_32delGCT [p.Leu11del]). Previously reported indels were detected in *CHEK2* and *PALB2*. In addition, five nonsense mutations, which have been previously reported by others, were found in six different patients (Table 5.3; details in Appendix D.14).

5.3.2.2 Missense Variants

Of the 155 unique missense variants (Appendix D.15), 119 were prioritized by consulting published literature, disease- and gene-specific databases. All are of unknown clinical significance and 21 have not been previously reported.

Missense variants that have been previously described as detrimental include the *ATM* variant c.7271T>G (p.Val2424Gly; rs28904921; two patients), which replaces a hydrophobic residue by glycine in the conserved FAT domain and confers a ninefold increase (95% CI) in BC risk [Goldgar et al., 2011]. Functional studies, assessing ATM kinase activity *in vitro* with TP53 as a substrate, showed that cell lines heterozygous for the mutation had less than 10% of wild-type kinase activity, such that this variant is expected to act in a dominant-negative manner [Chenevix-Trench et al., 2002]. The *CHEK2* variant c.433C>T (p.Arg145Trp; rs137853007; one patient) results in rapid degradation of the mutant protein [Lee et al., 2001]. Finally, the *PMS2* variant c.2T>C (p.Met1Thr) is listed in ClinVar as pathogenic and would be expected to abrogate correct

Table 5.3: Variants Resulting in Premature Protein Truncation

Gene	Exon	Variant	rsID (dbSNP142) Allele Frequency (%)[a]	Details
Frameshift Insertions/Deletions				
BRCA1	10 of 23	NM_007294.2:c.3550_3551insA[b] p.Gly1184Glufs	Novel	STOP at p.1187 676 AA short
PALB2	4 of 13	NM_024675.3:c.757_758delCT[c] p.Leu253Ilefs	rs180177092	STOP at p.255 932 AA short
PALB2	9 of 13	NM_024675.3:c.2920_2921delAA[c] p.Lys974Glufs	rs180177126	STOP at p.979 208AA short
Insertions/Deletions with Conserved Reading Frame				
CDH1	1 of 16	NM_004360.3:c.30_32delGCT[d] p.Leu11del	Novel	Loss of one AA Frame and AA sequence conserved
CHEK2	4 of 14	NM_007194.3:c.483_485delAGA[c] p.Glu161del	-	Loss of one AA Frame and AA sequence conserved
Stop Codons				

ATM	13 of 63	NM_000051.3:c.1924G>T[c] p.Glu642Ter	-	2415 AA short
ATM	62 of 63	NM_000051.3:c.8977C>T[c] p.Arg2993Ter	-	64 AA short
BRCA1	23 of 23	NM_007294.2:c.5503C>T[b] p.Arg1835Ter	rs41293465	28 AA short
PALB2	13 of 13	NM_024675.3: c.3549C>G[c] p.Tyr1183Ter	rs118203998	4 AA short

[a] If known.

[b] Not confirmed through Sanger sequencing.

[c] Confirmed by Sanger sequencing.

[d] Ambiguous Sanger sequencing results.

AA: amino acid.

initiation of translation. This variant has not been reported in BC families, but is associated with colorectal cancer (CRC) [Senter et al., 2008].

5.3.3 Variant Prioritization

We prioritized an average of 18.2 variants in each gene, ranging from seven (*XRCC2*) to 61 (*ATM*), an average of 0.41 variants/kb, and an average of 0.65 variants/patient (Table 5.4). *ATM* had the second greatest gene probe coverage (103,511 nt captured), the highest number of unique prioritized variants, and was among the top genes for number of prioritized variants per kilobase (0.59).

In total, our framework allowed for the prioritization of 346 unique variants in 246 patients, such that 85.7% of tested patients (N = 287) had at least one prioritized variant. Most patients (84.7%) harbored fewer than four prioritized variants. The distribution of patients with prioritized variants was similar across eligibility groups (Table 5.5). Although Class 5 (91.1% of patients with prioritized variants) and Class 8 (100% with prioritized variants, with a single patient in this category) deviated to a greater extent from the mean variants/category, these differences were not significant, χ^2 (4, N = 246) = 0.98, $p > 0.90$. The distribution of prioritized variants among mutation types is nine protein truncating, 28 mRNA splicing, 34 affecting RBBS and/or UTR structure, 36 affecting TFBS, 119 missense, and 149 affecting SRFBS, of which 29 were prioritized into multiple categories (Appendix D.16 and D.17 show this information by gene and patient, respectively).

All prioritized protein-truncating (N = 10), and selected splicing (N = 7) and missense (N = 5) variants were verified by bidirectional Sanger sequencing as they were more likely to be pathogenic (taking into account available published studies). Of the protein-truncating variants, four nonsense, one indel with a conserved reading frame, and two frameshifts were confirmed (Table 5.3). Six splicing variants and all missense were confirmed (Table 5.1 and Appendix D.15). An additional 145 prioritized variants, including 88 noncoding variants, were confirmed upon resequencing of patient gDNA. Of the 57 resequenced coding variants, 13 were prioritized for their noncoding effects (12 SRFBS, two cryptic site strengthening; one variant prioritized for both). These variants

Table 5.4: Comparing Counts of Prioritized Variants

Gene	Unique prioritized variants	Unique patients	Gene probe coverage (nt)	Prioritized variants/patient	Prioritized variants/kb
ATM	61	102	103511	0.60	0.59
ATP8B1	21	37	94793	0.57	0.22
BARD1	17	46	73735	0.37	0.23
BRCA1	19	24	52075	0.79	0.36
BRCA2	24	28	73332	0.86	0.33
CDH1	21	32	61312	0.66	0.34
CHEK2	12	13	28372	0.92	0.42
MLH1	18	25	50553	0.72	0.36
MRE11A	17	31	64713	0.55	0.26
MSH2	18	17	112437	1.06	0.16
MSH6	19	23	25216	0.83	0.75
MUTYH	8	16	21439	0.50	0.37
NBN	11	21	57067	0.52	0.19

PALB2	26	46	25319	0.57	1.03
PMS2[a]	8	15	11726	0.53	0.68
PTEN[b]	15	23	86059	0.65	0.17
RAD51B[c]	22	47	62465	0.47	0.35
STK11	12	20	28373	0.60	0.42
TP53	11	30	23544	0.37	0.47
XRCC2	7	10	19942	0.70	0.35

[a] Homologous to other genomic regions, thus fewer probes designed within gene.

[b] PTENhas pseudogene PTENP1, thus fewer probes covering exonic regions.

[c] Probes limited to 1,000 nt surrounding all exons, and 10,000 nt up- and downstream of gene.

Table 5.5: Distribution of Recruited Patients among Eligibility Groups

Eligibility Group[a]	Number of Patients within Eligibility Group	Number of Patients with Prioritized Variants
Breast cancer <60 year, and a first or second-degree relative with ovarian cancer or male breast cancer (5).	68	62
Breast and ovarian cancer in the same individual, or bilateral breast cancer with the first case <50 years (6).	37	32
Two cases of ovarian cancer, both <50 years, in first or second-degree relatives (7).	72	59
Two cases of ovarian cancer, any age, in first or second-degree relatives (8).	1	1
Three or more cases of breast or ovarian cancer at any age (10).	109	92
	287	246

The risk categories for individuals eligible for screening for a genetic susceptibility to breast or ovarian cancers are determined by the Ontario Ministry of Health and Long Term-Care Referral Criteria for Genetic Counseling.

[a] Numbers in parentheses correspond to eligibility group designation.

can be found in Appendix D.17 (where “coverage” column contains two or more coverage values).

5.3.3.1 Negative Control

ATP8B1 was sequenced and analyzed in all patients as a negative control (Appendix D.18). We prioritized 21 *ATP8B1* variants with an average of 0.22 variants/kb and 0.57 variants/patient. This is lower than the prioritization rate for many of the documented

HBOC genes. This result illustrates that the proposed method represents a screening rather than a diagnostic approach, as some variants may be incorrectly prioritized.

5.3.4 Pedigree Analysis

Pathogenic *BRCA2* variants within a region of exon 11 have been associated with a high incidence of OC. We therefore verified whether there were a high number of OC cases in the families of patients prioritized with exon 11 *BRCA2* variants (N = 3). The family of the patient with *BRCA2* variant c.4828G>A (p.Val1610Met; diagnosed with BC at 65) has three reported cases of BC/OC, one of which is OC (diagnosed at 74), two degrees of separation from the proband. The patient with *BRCA2*:c.6317T>C (p.Leu2106Pro; diagnosed with BC at 52) has three other affected family members, two with OC and one with BC. Finally, four patients found to have the *BRCA2* variant c.5199C>T (p.Ser1733 =) do not have any family members with reported cases of OC.

We also selected patients with prioritized mismatch repair variants (N = 8 in 10 patients) to assess the incidence of reported CRC cases in these families. Notably, the patient with mutation *MSH2*:c.1748A>G (p.Asn583Ser) had five relatives with CRC. A similar analysis of prioritized *CDH1* variants did not reveal any patients with a family history of gastric cancer.

5.3.5 LR Analyses

We carried out co-segregation analysis of 25 patients with prioritized pathogenic variants (four nonsense, four frameshift, two in-frame deletions, six missense, four natural splicing, and six cryptic splicing; including a patient who exhibited prioritized natural and cryptic SS variants). We compared these findings with those from patients (N = 25) harboring moderate-priority variants (variants prioritized through IT analysis only) and those in whom no variants were flagged or prioritized (N = 14). In instances where disease alleles could be transmitted through either founder parent, the lineage with the highest likelihood ratio (LR) was reported. For patients with likely pathogenic variants, the LRs ranged from 0.00 to 70.96 (Table 5.6 and Appendix D.19). Disease co-segregation was supported (LR > 1.0) in 18 patients, and the remainder were either neutral (LR < 1.0 [Mohammadi et al., 2009]) or could not be analyzed either due to

Table 5.6: LR Values for Patients with Prioritized Truncating, Splicing, and Selected Missense Variants

Genes[a]	Variant		Category	UWO ID	LR
	mRNA	Protein			
ATM	c.1924G>T	p.Glu642Ter	Nonsense	10-2F	7.46^{MGM},9.61^{MGF}
	c.6198+1G>A	-	Natural splicing	8-1D.9-1B	1.00
	c.7271T>G	p.Val2424Gly	Missense	10-1F	1.44
				12-1D	1.96^{P}
	c.8977C>T	p.Arg2993Ter	Nonsense	12-4G.13-5D	5.30^{P}
BARD1	c.1454C>T	p.Ala485Val	Cryptic splicing	8-1D.9-1B	1.00
BRCA1	c.3550_3551insA	p.Gly1184Glufs	Frameshift indel	11-6H	3.36^{P}
	c.5503C>T	p.Arg1835Ter	Nonsense	8-5D.9-5D	41.99
BRCA2	c.10095delCins11	p.Ser3366Asnfs	Frameshift indel	15-4E	3.71
CDH1	c.30_32delGCT	p.Leu11del	Inframe deletion	10-4A	1.00
	c.1223C>G	p.Ala408Gly	Cryptic splicing	15-3G	2.14
CHEK2	c.470T>C	p.Ile157Thr	Missense	12-2G	2.86
				15-5G	19.44^{P}

	c.433C>T	p.Arg145Trp	Missense	4-3C.5-4G.14-4A	3.48
PALB2	c.3549C>G	p.Tyr1183Ter	Nonsense	15-1E	1.78
	c.757_758delCT	p.Leu253Ilefs	Frameshift indel	10-6F	70.96
	c.2920_2921delAA	p.Lys974Glufs	Frameshift indel	8-3A.9-3A	5.03
PMS2	c.2T>C	p.Met1Thr	Missense	11-4H	16.53^{P}
RAD51B	c.84G>A	p.Gln28=	Leaky splicing	8-1H.9-1E	3.51^{P}
	c.958-29A>T	-	Cryptic splicing	10-4B	7.44^{P}
STK11	c.375-194GT>AC	-	Cryptic splicing	10-5A	2.67^{M}

LR values in favor of neutrality are not shown.

aRefSeq accessions: *ATM* (NM_000051.3), *BARD1* (NM_000465.2), *BRCA1* (NM_007294.2), *BRCA2* (NM_000059.3), *CDH1* (NM_004360.3), *CHEK2* (NM_007194.3), *PALB2* (NM_024675.3), *PMS2* (NM_000535.5), *RAD51B* (NM_002877.4), *STK11* (NM_000455.4).

P, paternal; M, maternal; MGF, maternal grandfather; MGM, maternal grandmother.

missing pedigree information or limited numbers of affected individuals in a family. Patient 10–6F (*PALB2*: c.757 758delCT) exhibited the highest likelihood (LR = 70.96). Prioritized variants with neutral evidence include a variant that abolishes a natural SS in *MRE11A*, c.2070+2T>A (LR = 0.03), and an in-frame deletion c.483 485delAGA in *CHEK2* (LR = 0.00).

5.4 Discussion

Rare noncoding and/or non-truncating mutations can confer an increased risk of disease in BC [Tavtigian et al., 2009]. This study determined both coding and noncoding sequences of 20 HBOC-related genes, with the goal of discovering and prioritizing rare variants with potential effects on gene expression. This work emphasizes results from the analysis of noncoding variants, which are abundant in these genes, yet have been underrepresented in previous HBOC mutation analyses. Nevertheless, alterations to mRNA binding sites in BRCA, and lower risk or rare HBOC genes, have been shown to contribute to HBOC (exonic splicing enhancers (ESEs) in *ATM* [Heikkinen et al., 2005], *BARD1* [Ratajska et al., 2011], and *BRCA* genes [Gochhait et al., 2007; Sanz et al., 2010]). We prioritized 346 unique variants that were predicted to result in four nonsense, three frameshift, two indels with preserved reading frame, 119 missense, four natural splicing, six cryptic splicing, 17 pseudoexon activating, 148 SRFBS, 36 TFBS, three UTR structure, and 31 RBBS mutations (Appendix D.16). Among these variants, 101 were novel (see Appendix D.20 for references to previously identified variants). Compared to our initial seven-gene panel [Mucaki et al., 2016], the inclusion of the additional genes in this study prioritized at least one variant in 15% additional patients (increased from 70.6 to 85.7%).

The *BRCA* genes harbor the majority of known germline pathogenic variants for HBOC families [Chong et al., 2014]. However, a large proportion of the potentially pathogenic variants identified in our study were detected in *ATM*, *PALB2*, and *CHEK2*, which, although of lower penetrance, were enriched because the eligibility criteria excluded known *BRCA1* and *BRCA2* carriers. *BRCA1* and *BRCA2* variants were nevertheless prioritized in some individuals. We also had expected intragenic clustering of some *BRCA* coding variants [Mucaki et al., 2011]. For example, pathogenic variants occurring

within exon 11 of *BRCA2* are known to be associated with higher rates of OC in their families [Lubinski et al., 2004]. We identified three variants in exon 11; however, there was no evidence of OC in these families. Overall, *ATM* and *PALB2* had the highest number of prioritized variants (61 and 26, respectively). However, only 12 variants were prioritized in *CHEK2*; potentially pathogenic variants may have been underrepresented during sequence alignment as a consequence of the known paralogy with *CHEK2P2*.

Fewer *TP53*, *STK11*, and *PTEN* variants were prioritized, as pathogenic variants in these genes tend to be infrequent in patients who do not fulfill the clinical criteria for their associated syndromes (Li-Fraumeni syndrome, Peutz-Jeghers syndrome, and Cowden syndrome, respectively [Hollestelle et al., 2010]), although they have been indicated as near moderate to high-risk genes in BC [Easton et al., 2015]. This underrepresentation of prioritized variants may be supported by the negative Residual Variation Intolerance Scores (RVIS) for these genes [Petrovski et al., 2013], which are likely indicative of purifying selection. Although the density of prioritized variants in these genes is below average (18.2 per gene), the total number was nonetheless important (*TP53* = 11, *STK11* = 12, *PTEN* = 15).

The fundamental difference between IT and other approaches such as combined annotation-dependent depletion [CADD; Kircher et al., 2014] is that IT depends only on positive experimental data from the same or closely related species. CADD does not appear to account for unobserved reversions or other hidden mutations [e.g. perform a Jukes–Cantor correction; Jukes and Cantor, 1969], nor are the effects of these simulated. Furthermore, the CADD scoring system is *ad hoc*, which contrasts with strong theoretical basis on the IT approach [Rogan and Schneider, 1995] in which information changes in bits represent a formally proven relationship to thermodynamic stability, and therefore can be used to measure binding affinity. This makes it different from other unitless methods with unknown distributions, in which differences in binding affinity cannot be accurately extrapolated from derived scores.

We compared the frequency of all prioritized variants in our patient cohort to the population allele frequencies [1000 Genomes Project, Phase 3;

http://www.1000genomes.org; 1000 Genomes Project Consortium, 2012] to determine if variants more common in our cohort might be suggestive of HBOC association. Three variants in at least five HBOC patients are present at a much lower frequency in the general population than in our HBOC population. *NBN* c.∗2129G>T, present in 4.18% of study cohort, is considerably rarer globally (0.38% in 1000 Genomes; <0.1% in other populations). Similarly, the *RAD51B* c.-3077G>T variant (2.09%) is rare in the general population (0.08%). Interestingly, *BARD1* c.33G>T (1.74% of study cohort) has only been reported in the American and European populations in 1000 Genomes (0.29% and 0.20%, respectively) and only Europeans in the Exome Variant Server (0.24%; http://evs.gs.washington.edu/EVS/). In Southwestern Ontario, individuals are often of American or European ancestry. The variant was found to be more common in the Exome Aggregation Consortium (ExAC; http://exac.broadinstitute.org/) in 1.17% tested Finnish population (0.41% in their non-Finnish European cohort), though no alleles were found in the Finnish populations in 1000 Genomes (N = 99). Therefore, the allele frequency of this *BARD1* variant in our HBOC population may simply be enriched in a founder subset of general populations. While we cannot rule out skewing of these allele frequencies due to population stratification, our findings suggest that gene expression levels could be impacted by these variants.

We applied subpopulation allele frequency analysis for all of our prioritized variants. Appendix D.21 lists the 49 variants that have allele frequencies >1% in various subpopulations (based on dbSNP). Allele frequencies were as high as 4.2% for the *BRCA2* c.-40+192C>T (8-1G.9-1C), predicted to affect TF binding, in the East Asian subpopulation. Without additional information on patient ethnicities, it is not possible to eliminate prioritized variants that are common in specific subpopulations.

Co-segregation analysis is recommended by the American College of Medical Genetics and Genomics for variant classification [Richards et al., 2015]. Among patients with likely pathogenic, highly penetrant mutations in our cohort (N = 24), some variants had LR values consistent with causality, whereas others provided little evidence to support co-segregation among family members (Table 5.6 and Appendix D.19). An important

caveat, however, was that the use of *BRCA2* penetrance values in non-*BRCA* genes may have resulted in underestimates of LR values.

In order to evaluate the application of co-segregation analysis in the context of this study, we chose to perform this analysis on patients with moderate priority variants (i.e., variants affecting binding sites) and patients with no flagged or prioritized variants (N = 25 and 14, respectively). LRs ranged from 0.0034 to 78.0 for moderate-priority variants and from 0.0005 to 57.0 for patients with no flagged or prioritized variants (Appendix D.5-Supp. Fig. S2). The proportion of LR values supporting neutrality and those supporting causation was comparable between patients with prioritized, moderately prioritized, and flagged variants (Appendix D.5-Supp. Fig. S2). This suggests that co-segregation analysis is only useful in the context of other supporting results for assessing pathogenicity (e.g., likelihood of being pathogenic or benign). Furthermore, the lack of genotype information and at times smaller pedigrees likely also contributed to the lack of concordance between LRs and variant priority.

A small number of patients with a known pathogenic variant carried other prioritized variants. These were likely benign or possibly phenotypic modifiers. One patient possessed five prioritized variants (one missense, one SRFBS, one TFBS, and two RBBSs) in addition to a *BRCA1* nonsense mutation (c.5503C>T). While these variants may not directly contribute to causing HBOC, they may act as a risk modifier and alter expression levels [Antoniou and Easton, 2006].

Similarly, genes lacking association with HBOC can be used as a metric for determining a false-positive rate of variant prioritization. In this study, we prioritized 21 *ATP8B1* variants among 37 of our HBOC patients (Appendix D.18) despite it having not been previously associated with any type of cancer. A variant with a deleterious effect on *ATP8B1* may lead to *ATP8B1*-related diseases, such as progressive familial intrahepatic cholestasis [Gonzales et al., 2014], but should not increase the chances of developing BC. Thus, while our framework may be effective at prioritizing variants, only genes with previous association to a disease should be included in analyses similar to the present study to minimize falsely prioritized variants.

Additional workup of prioritized noncoding and non-*BRCA* variants is particularly important, because with few exceptions [Easton et al., 2015], the pathogenicity of many of the genes and variants has not been firmly established. Furthermore, mutations in several of these genes confer risk to other types of cancer, which alters the management of these patients [Knappskog and Lønning, 2012]. The next step toward understanding the role these prioritized variants play in HBOC is to test family members of the proband and to carry out functional analysis. If this is not possible, then their effects on gene expression could be evaluated using assays for RNA stability and RNA localization. Protein function could be evaluated by binding site assays, protein activity, and quantitative PCR.

A significant challenge associated with VUS analysis, particularly in the case of many of these recent HBOC gene candidates, is the underreporting of variants and thus positive findings tend to be overrepresented in the literature [Kraft, 2008]. Hollestelle et al. [2010] argue that a more stringent statistical standard must be applied (i.e., *P*-values of 0.01 should be used as opposed to 0.05) to underreported variants (namely in moderate-risk alleles), because of failure to replicate pathogenic variants, which we have also found [Viner et al., 2014]. In the same way that we use IT-based analysis to justify prioritizing variants for further investigation, variants that are disregarded as lower priority (and that are likely not disease causing) have been subjected to the same thresholds and criteria. Integrating this set of labeled prioritized and flagged, often rare variants from this cohort of *BRCA*-negative HBOC patients, to findings from exome or gene panel studies of HBOC families should accelerate the classification of some VUS.

Different variant interpretation and reporting guidelines consider the reporting of VUS to be either optional or essential [Wallis et al., 2013; Richards et al., 2015]. In all cases, a reported VUS cannot be the basis for a clinical decision and should be followed up and further investigated. In any case, the number of reported VUS in an individual is frequently too extensive for detailed characterization. Reducing the full set of variants obtained by complete gene sequencing to a prioritized list will be an essential prerequisite for targeting potentially clinically relevant information. Informing patients of prioritized VUS may increase patient accrual and participation [Murphy et al., 2008]. However, it

will be critical to explain both the implications and significance of prioritization and the limitations, namely counselling patients to avoid clinical decisions, based on this information [Vos et al., 2012].

5.5 References

1000 Genomes Project Consortium, 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. 491:56–65.

Al Bakir M, Gabra H. 2014. The molecular genetics of hereditary and sporadic ovarian cancer: implications for the future. Br Med Bull. 112:57–69.

Antoniou AC, Easton DF. 2006. Models of genetic susceptibility to breast cancer. Oncogene. 25:5898–5905.

Antoniou A, Pharoah PDP, Narod S, Risch HA, Eyfjord JE, Hopper JL, Loman N, Olsson H, Johannsson O, Borg A, Pasini B, Radice P, et al. 2003. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. Am J Hum Genet. 72:1117–1130.

Apostolou P, Fostira F. 2013. Hereditary breast cancer: the era of new susceptibility genes. BioMed Res Int. 2013:e747318.

Biswas DK, Shi Q, Baily S, Strickland I, Ghosh S, Pardee AB, Iglehart JD. 2004. NF-kappaB activation in human breast cancer specimens and its role in cell proliferation and apoptosis. Proc Natl Acad Sci USA. 101:10137–10142.

Caminsky NG, Mucaki EJ, Rogan PK. 2015. Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information theoretical analysis. F1000Res. 3:282.

Cancer Genome Atlas Network. 2012. Comprehensive molecular portraits of human breast tumours. Nature. 490:61–70.

Cassa CA, Savage SK, Taylor PL, Green RC, McGuire AL, Mandl KD. 2012. Disclosing pathogenic genetic variants to research participants: quantifying an emerging ethical responsibility. Genome Res. 22:421–428.

Chenevix-Trench G, Spurdle AB, Gatei M, Kelly H, Marsh A, Chen X, Donn K, Cummings M, Nyholt D, Jenkins MA, Scott C, Pupo GM, et al. 2002. Dominant negative ATM mutations in breast cancer families. J Natl Cancer Inst. 94:205–215.

Chong HK, Wang T, Lu H-M, Seidler S, Lu H, Keiles S, Chao EC, Stuenkel AJ, Li X, Elliott A M. 2014. The validation and clinical implementation of BRCAplus: a comprehensive high -risk breast cancer diagnostic assay. PloS One. 9:e97408.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Angel G del, Rivas MA, Hanna M, McKenna A, Fennell TJ, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43:491–498.

Ding H, Mao C, Li S-M, Liu Q, Lin L, Chen Q. 2011. Lack of association between ATM C.1066-6T>G mutation and breast cancer risk: a meta-analysis of 8,831 cases and 4,957 controls. Breast Cancer Res Treat. 125:473–477.

Dorman SN, Shirley BC, Knoll JHM, Rogan PK. 2013. Expanding probe repertoire and improving reproducibility in human genomic hybridization. Nucleic Acids Res. 41:e81.

Dorman S, Viner C, Rogan P. 2014. Splicing mutation analysis reveals previously unrecognized pathways in lymph node-invasive breast cancer. Sci Rep. 4:7063.

Drost M, Koppejan H, Wind N de. 2013. Inactivation of DNA mismatch repair by variants of uncertain significance in the PMS2 gene. Hum Mutat. 34:1477–1480.

Duzkale H, Shen J, McLaughlin H, Alfares A, Kelly MA, Pugh TJ, Funke BH, Rehm HL, Lebo MS. 2013. A systematic approach to assessing the clinical significance of genetic variants. Clin Genet. 84:453–463.

Easton DF, Pharoah PDP, Antoniou AC, Tischkowitz M, Tavtigian SV, Nathanson KL, Devilee P, Meindl A, Couch FJ, Southey M, Goldgar DE, Evans DGR, et al. 2015. Gene-panel sequencing and the prediction of breast-cancer risk. N Engl J Med. 372:2243–2257.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature. 489:57–74.

Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, Bishop DT, Weber B, Lenoir G, Chang-Claude J, Sobol H, Teare MD, et al. 1998. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. Am J Hum Genet. 62:676–689.

Gadiraju S, Vyhlidal CA, Leeder JS, Rogan PK. 2003. Genome-wide prediction, display and refinement of binding sites with information theory-based models. BMC Bioinformatics. 4:38.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, et al. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 27:182–189.

Gochhait S, Bukhari SIA, Bairwa N, Vadhera S, Darvishi K, Raish M, Gupta P, Husain SA, Bamezai RNK. 2007. Implication of BRCA2 -26G>A 5’ untranslated region polymorphism in susceptibility to sporadic breast cancer and its modulation by p53 codon 72 Arg>Pro polymorphism. Breast Cancer Res. 9:R71.

Goldgar DE, Healey S, Dowty JG, Da Silva L, Chen X, Spurdle AB, Terry MB, Daly MJ, Buys SM, Southey MC, Andrulis I, John EM, et al. 2011. Rare variants in the ATM gene and risk of breast cancer. Breast Cancer Res. 13:R73.

Gonzales E, Spraul A, Jacquemin E. 2014. Clinical utility gene card for: progressive familial intrahepatic cholestasis type 1. Eur J Hum Genet. 22(4). doi:10.1038/ejhg.2013.188.

Halvorsen M, Martin JS, Broadaway S, Laederach A. 2010. Disease-associated mutations that alter the RNA structural ensemble. PLoS Genet. 6:e1001074.

Heikkinen K, Karppinen S-M, Soini Y, Makinen M, Winqvist R. 2003. Mutation screening of Mre11 complex genes: indication of RAD50 involvement in breast and ovarian cancer susceptibility. J Med Genet. 40:e131.

Heikkinen K, Rapakko K, Karppinen S-M, Erkko H, Nieminen P, Winqvist R. 2005. Association of common ATM polymorphism with bilateral breast cancer. Int J Cancer. 116:69–72.

Hollestelle A, Wasielewski M, Martens JW, Schutte M. 2010. Discovering moderate-risk breast cancer susceptibility genes. Curr Opin Genet Dev. 20:268–276.

Howlander N, Noone AM, Krapcho M, Garshell J, Miller D, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). 2014. SEER Cancer Statistics Review 1975-2011, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975 2011/, based on November 2013 SEER data submission, posted on the SEER web site April 2014.

Janatova M, Soukupova J, Stribrna J, Kleiblova P, Vocka M, Boudova P, Kleibl Z, Pohlreich P. 2015. Mutation analysis of the RAD51C and RAD51D genes in high-risk ovarian cancer patients and families from the Czech Republic. PloS One 10:e0127711.

Jones S, Emmerson P, Maynard J, Best JM, Jordan S, Williams GT, Sampson JR, Cheadle JP. 2002. Biallelic germline mutations in MYH predispose to multiple colorectal adenoma and somatic G:C→T:A mutations. Hum Mol Genet 11:2961–2967.

Jukes T, Cantor C. 1969. Evolution of protein molecules. (Munro H N, ed.). Mammalian protein metabolism III, New York: Academic Press, p 21–132.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46: 310–315.

Knappskog S, Lønning PE. 2012. P53 and its molecular basis to chemoresistance in breast cancer. Expert Opin. Ther. Targets 16 Suppl 1: S23–30.

von Kodolitsch Y, Berger J, Rogan PK. 2006. Predicting severity of haemophilia A and B splicing mutations by information analysis. Haemophilia 12:258–262.

Kraft P. 2008. Curses--winner's and otherwise--in genetic epidemiology. Epidemiol. Camb. Mass 19: 649–651; discussion 657–658.

Lee SB, Kim SH, Bell DW, Wahrer DC, Schiripo TA, Jorczak MM, Sgroi DC, Garber JE, Li FP, Nichols KE, Varley JM, Godwin AK, et al. 2001. Destabilization of CHK2 by a missense mutation associated with Li-Fraumeni Syndrome. Cancer Res. 61: 8062–8067.

Ligtenberg MJL, Kuiper RP, Chan TL, Goossens M, Hebeda KM, Voorendt M, Lee TYH, Bodmer D, Hoenselaar E, Hendriks-Cornelissen SJB, Tsui WY, Kong CK, et al. 2009. Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1. Nat. Genet. 41: 112–117.

Lu R, Mucaki EJ, Rogan PK. 2017. Discovery and validation of information theory-based transcription factor and cofactor binding site motifs, Nucleic Acids Research, 45(5): e27.

Lubinski J, Phelan CM, Ghadirian P, Lynch HT, Garber J, Weber B, Tung N, Horsman D, Isaacs C, Monteiro ANA, Sun P, Narod SA. 2004. Cancer variation associated with the position of the mutation in the BRCA2 gene. Fam. Cancer 3: 1–10.

Maxwell KN, Domchek SM. 2013. Familial Breast Cancer Risk. Curr. Breast Cancer Rep. 5: 170–182.

Minion LE, Dolinsky JS, Chase DM, Dunlop CL, Chao EC, Monk BJ. 2015. Hereditary predisposition to ovarian cancer, looking beyond BRCA1/BRCA2. Gynecol. Oncol. 137: 86–92.

Mohammadi L, Vreeswijk MP, Oldenburg R, Ouweland A van den, Oosterwijk JC, Hout AH van der, Hoogerbrugge N, Ligtenberg M, Ausems MG, Luijt RB van der,

Dommering CJ, Gille JJ, et al. 2009. A simple method for co-segregation analysis to evaluate the pathogenicity of unclassified variants; BRCA1 and BRCA2 as an example. BMC Cancer 9: 211.

Mucaki EJ, Ainsworth P, Rogan PK. 2011. Comprehensive prediction of mRNA splicing effects of BRCA1 and BRCA2 variants. Hum. Mutat. 32: 735–742.

Mucaki EJ, Caminsky NG, Perri AM, Lu R, Laederach A, Halvorsen M, Knoll JHM, Rogan PK. 2016. A unified analytic framework for prioritization of non-coding variants of uncertain significance in heritable breast and ovarian cancer. BMC Med. Genomics. 9: 19.

Mucaki EJ, Shirley BC, Rogan PK. 2013. Prediction of Mutant mRNA Splice Isoforms by Information Theory-Based Exon Definition. Hum. Mutat. 34: 557–565.

Murphy J, Scott J, Kaufman D, Geller G, LeRoy L, Hudson K. 2008. Public expectations for return of results from large-cohort genetic research. Am. J. Bioeth. AJOB 8: 36–43.

Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. 2013. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. 9: e1003709.

Ratajska M, Antoszewska E, Piskorz A, Brozek I, Borg Å, Kusmierek H, Biernat W, Limon J. 2011. Cancer predisposing BARD1 mutations in breast–ovarian cancer families. Breast Cancer Res. Treat. 131: 89–97.

Reiman A, Srinivasan V, Barone G, Last JI, Wootton LL, Davies EG, Verhagen MM, Willemsen MA, Weemaes CM, Byrd PJ, Izatt L, Easton DF, et al. 2011. Lymphoid tumours and breast cancer in ataxia telangiectasia; substantial protective effect of residual ATM kinase activity against childhood tumours. Br. J. Cancer 105: 586–591.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL, et al. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American

College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. Off. J. Am. Coll. Med. Genet. 17: 405–424.

Rogan PK, Faux BM, Schneider TD. 1998. Information analysis of human splice site mutations. Hum. Mutat. 12: 153–171.

Rogan PK, Schneider TD. 1995. Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites. Hum Mutat 6:74–76.

Rogan PK, Svojanovsky S, Leeder JS. 2003. Information theory-based analysis of CYP2C19, CYP2D6 and CYP3A5 splicing mutations. Pharmacogenetics 13:207–218.

Rogan PK, Zou GY. 2013. Best practices for evaluating mutation prediction methods. Hum Mutat 34:1581–1582

Santagata S, Hu R, Lin NU, Mendillo ML, Collins LC, Hankinson SE, Schnitt SJ, Whitesell L, Tamimi RM, Lindquist S, Ince TA. 2011. High levels of nuclear heat-shock factor 1 (HSF1) are associated with poor prognosis in breast cancer. Proc. Natl. Acad. Sci. U. S. A. 108: 18378–18383.

Santos C, Peixoto A, Rocha P, Pinto P, Bizarro S, Pinheiro M, Pinto C, Henrique R, Teixeira MR. 2014. Pathogenicity Evaluation of BRCA1 and BRCA2 Unclassified Variants Identified in Portuguese Breast/Ovarian Cancer Families. J. Mol. Diagn. 16: 324–334.

Sanz DJ, Acedo A, Infante M, Durán M, Pérez-Cabornero L, Esteban-Cardeñosa E, Lastra E, Pagani F, Miner C, Velasco EA. 2010. A High Proportion of DNA Variants of BRCA1 and BRCA2 Is Associated with Aberrant Splicing in Breast/Ovarian Cancer Patients. Clin. Cancer Res. 16: 1957–1967.

Schneider TD. 1997. Information content of individual genetic sequences. J Theor Biol 189:427–441.

Schrader KA, Masciari S, Boyd N, Salamanca C, Senz J, Saunders DN, Yorida E, Maines-Bandiera S, Kaurah P, Tung N, Robson ME, Ryan PD, et al. 2011. Germline mutations in CDH1 are infrequent in women with early-onset or familial lobular breast cancers. J. Med. Genet. 48: 64–68.

Seal S, Thompson D, Renwick A, Elliott A, Kelly P, Barfoot R, Chagtai T, Jayatilake H, Ahmed M, Spanova K, North B, McGuffog L, et al. 2006. Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. Nat. Genet. 38: 1239–1241.

Senter L, Clendenning M, Sotamaa K, Hampel H, Green J, Potter JD, Lindblom A, Lagerstedt K, Thibodeau SN, Lindor NM, Young J, Winship I, et al. 2008. The clinical phenotype of Lynch syndrome due to germ-line PMS2 mutations. Gastroenterology 135: 419–428.

Shirley BC, Mucaki EJ, Whitehead T, Costea PI, Akan P, Rogan PK. 2013. Interpretation, stratification and evidence for sequence variants affecting mRNA splicing in complete human genome sequences. Genomics Proteomics Bioinformatics 11:77–85.

Stankovic T, Kidd AM, Sutcliffe A, McGuire GM, Robinson P, Weber P, Bedenham T, Bradwell AR, Easton DF, Lennox GG, Haites N, Byrd PJ, et al. 1998. ATM mutations and phenotypes in ataxia-telangiectasia families in the British Isles: expression of mutant ATM and the risk of leukemia, lymphoma, and breast cancer. Am. J. Hum. Genet. 62: 334–345.

Stratton JF, Pharoah P, Smith SK, Easton D, Ponder BA. 1998. A systematic review and meta-analysis of family history and risk of ovarian cancer. Br. J. Obstet. Gynaecol. 105: 493–499.

Tavtigian SV, Oefner PJ, Babikyan D, Hartmann A, Healey S, Le Calvez-Kelm F, Lesueur F, Byrnes GB, Chuang S-C, Forey N, Feuchtinger C, Gioia L, et al. 2009. Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer. Am. J. Hum. Genet. 85: 427–446.

Théry JC, Krieger S, Gaildrat P, Révillion F, Buisine M-P, Killian A, Duponchel C, Rousselin A, Vaur D, Peyrat J-P, Berthet P, Frébourg T, et al. 2011. Contribution of bioinformatics predictions and functional splicing assays to the interpretation of unclassified variants of the BRCA genes. Eur. J. Hum. Genet. EJHG 19: 1052–1058.

Tournier I, Vezain M, Martins A, Charbonnier F, Baert-Desurmont S, Olschwang S, Wang Q, Buisine MP, Soret J, Tazi J, Frébourg T, Tosi M. 2008. A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. Hum. Mutat. 29: 1412–1424.

Vidal LJ-P, Perry JK, Vouyovitch CM, Pandey V, Brunet-Dunand SE, Mertani HC, Liu D-X, Lobie PE. 2010. PAX5alpha enhances the epithelial behavior of human mammary carcinoma cells. Mol. Cancer Res. MCR 8: 444–456.

Vihinen M. 2013. Guidelines for reporting and using prediction tools for genetic variation analysis. Hum Mutat 34:275–282.

Viner C, Dorman SN, Shirley BC, Rogan PK. 2014. Validation of predicted mRNA splicing mutations using high-throughput transcriptome data. F1000Research 3: 8.

Vos J, Gómez-García E, Oosterwijk JC, Menko FH, Stoel RD, Asperen CJ van, Jansen AM, Stiggelbout AM, Tibben A. 2012. Opening the psychological black box in genetic counseling. The psychological impact of DNA testing is predicted by the counselees' perception, the medical impact by the pathogenic or uninformative BRCA1/2-result. Psychooncology. 21: 29–42.

Wallis Y, Payne S, McAnulty C, Bodmer D, Sistermans E, Robertson K, Moore D, Abbs S, Deans Z, Devereau A. 2013. Practice Guidelines for the Evaluation of Pathogenicity and the Reporting of Sequence Variants in Clinical Molecular Genetics. Assoc. Clin. Genet. Sci. ACGS Dutch Soc. Clin. Lab. Spec. VKGL.

Walsh T, Casadei S, Lee MK, Pennil CC, Nord AS, Thornton AM, Roeb W, Agnew KJ, Stray SM, Wickramanayake A, Norquist B, Pennington KP, et al. 2011. Mutations in 12

genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing. Proc. Natl. Acad. Sci. U. S. A. 108: 18032–18037.

Zahnow CA. 2009. CCAAT/enhancer-binding protein beta: its role in breast cancer and associations with receptor tyrosine kinases. Expert Rev. Mol. Med. 11: e12.

Chapter 6

6 Clustered, Information-dense Transcription Factor Binding Sites Identify Genes with Similar Tissue-wide Expression Profiles

The work presented in this chapter is reproduced from:

Lu,R., Rogan,P.K. (2018) Clustered, information-dense transcription factor binding sites identify genes with similar tissue-wide expression profiles. *bioRxiv*. doi: https://doi.org/10.1101/283267

6.1 Background

The distinctive organization and combination of transcription factor binding sites (TFBSs) and regulatory modules in promoters dictates specific expression patterns within a set of genes [1]. Clustering of multiple adjacent binding sites for the same TF (homotypic clusters) and for different TFs (heterotypic clusters) defines cis-regulatory modules (CRMs) in human gene promoters and can amplify the influence of individual TFBSs on gene expression through increasing binding affinities, facilitated diffusion mechanisms and funnel effects [2]. Because tissue-specific TF-TF interactions in TFBS clusters are prevalent, these features can assist in identifying correct target genes by altering binding specificities of individual TFs [3]. Previously, we derived information - theory-based position weight matrices (iPWMs) from ChIP-seq data that can accurately detect TFBSs and quantify their strengths by computing associated R_i values (Rate of Shannon information transmission for an individual sequence [4]), with $R_{sequence}$ being the average of R_i values of all binding site sequences and representing the average binding strength of the TF [3]. Furthermore, information density-based clustering (IDBC) can effectively identify functional TF clusters by taking into account both the spatial organization (i.e. intersite distances) and information density (i.e. R_i values) of TFBSs [5].

TF binding profiles, either derived from in vivo ChIP-seq peaks [6–8] or computationally detected binding sites and CRMs [9], have been shown to be predictive of absolute gene expression levels using a variety of tissue-specific machine learning classifiers and regression models. Because signal strengths of ChIP-seq peaks are not strictly proportional to TFBS strengths [3], representing TF binding strengths by ChIP-seq signals may not be appropriate; nevertheless, both achieved similar accuracy [10]. CRMs have been formed by combining two or three adjacent TFBSs [9], which is inflexible, as it arbitrarily limits the number of binding sites contained in a module, and does not consider differences between information densities of different CRMs. Chromatin structure (e.g. histone modification (HM) and DNase I hypersensitivity) were also found to be statistically redundant with TF binding in explaining tissue-specific mRNA transcript abundance at a genome-wide level [7,8,11,12], which was attributed to the heterogeneous distribution of HMs across chromatin domains [8]. Combining these two types of data explained the largest fraction of variance in gene expression levels in multiple cell lines [7,8], suggesting that either contributes unique information to gene expression that cannot be compensated for by the other.

The number of genes directly bound by a TF significantly exceeds the number of differentially expressed (DE) genes whose expression levels significantly change upon knockdown of the TF. Only a small subset of direct target genes whose promoters overlap ChIP-seq peaks were DE after individually knocking 59 TFs down using small interfering RNAs (siRNAs) in the GM19238 cell line [13]. Using these knockdown data on 8,872 genes as the gold standard, correlation between TFBS counts and gene expression levels across 10 different cell lines was more predictive of DE targets than setting a minimum threshold on TFBS counts [14]. Their TFBS counts were defined as the number of ChIP-seq peaks overlapping the promoter, though it was unknown how many binding sites were present in these peaks; true positives might not be direct targets in the TF regulatory cascade, as the promoters of these targets were not intersected with ChIP-seq peaks. By perturbing gene expression with CAS9-directed clustered regularly interspaced short palindromic repeats (CRISPR) of 10 different TF genes in K562 cells, the regulatory effects of each TF on 22,046 genes were dissected by single cell RNA sequencing with a regularized linear computational model [15]; this accurately revealed

DE targets and new functions of individual TFs, some of which were likely regulated through direct interactions at TFBS in their corresponding promoters. Machine learning classifiers have also been applied in a small number of gene instances to predict targets of a single TF using features extracted from n-grams derived from consensus binding sequences [16], or from TFBSs and homotypic binding site clusters [5].

To investigate whether the distribution and composition of CRMs in promoters substantially determines gene expression profiles of direct TF targets, we developed a general machine learning framework that predicts which genes have similar expression profiles to a given gene and predicts DE direct TF targets by combining information theory-based TF binding profiles with DHSs. Upon filtering for accessible promoter intervals with DHSs, features designed to capture the spatial distribution and information composition of CRMs were extracted from clusters identified by the IDBC algorithm from iPWM-detected TFBSs. Though not all direct targets regulated by multiple TFs share a common tissue-wide expression profile, this framework provides insight into the transcriptional program of genes with similar profiles by dissecting their cis-regulatory element organization and strengths. We identify genes with comparable tissue-wide expression profiles by application of Bray-Curtis similarity [17]. Using transcriptome data generated by CRISPR- [15] and siRNA-based [13] TF knockdowns, we predicted DE TF target genes that are simultaneously direct targets whose promoters overlap tissue-specific ChIP-seq peaks, in contrast with the correlation-based approach [14].

6.2 Methods

To identify genes with similar tissue-wide expression patterns, we formally define gene expression profiles and pairwise similarity measures between profiles of different genes. A general machine learning framework relates features extracted from the organization of TFBSs in these genes to their tissue-wide expression patterns. Positives and negatives for predicting DE direct protein-coding (PC) TF target genes (TF targets for short below) were obtained from CRISPR- and siRNA-generated knockdown data (see below). (The results on the siRNA-generated knockdown data are in progress.)

6.2.1 Similarity between Gene Expression Profiles

For each of 56,238 genes, the Genotype-Tissue Expression (GTEx) project measured its expression levels in 53 tissues in a number of individuals (*N*=5-564), and provides the median expression value (in RPKM (Reads Per Kilobase of transcript per Million mapped reads) in the GTEx Analysis v6p release) of each gene in each tissue [18]. To capture the tissue-wide overall expression pattern of a gene instead of within a single tissue, the expression profile of a gene was defined as its median RPKM across the 53 tissues (Equation 6.1), which forms a vector of size 53 and does not distinguish between different isoforms whose expression patterns may significantly differ from each other.

$$EP^A = [MEV_1^A, MEV_2^A, \dots, MEV_{53}^A] \ \ (in\ RPKM) \quad [6.1]$$

where EP^A is the expression profile of Gene A, MEV_1^A is the median expression value of Gene A in Tissue 1, MEV_2^A is the median expression value of Gene A in Tissue 2, etc.

To obtain ground-truth genes that have similar expression profiles to a given gene, the Bray-Curtis Similarity (Equation 6.2) was used to compute the similarity value between the expression profiles of two genes. Compared to other similarity metrics (Table 6.1, Example 6.1), its application is justified by three desired properties, including 1) maintaining bounds of 0 and 1, 2) achieving the maximal similarity 1 if and only if two vectors are identical, and 3) larger values having a larger impact on the resultant similarity value.

$$sim_{Bray-Curtis}(EP^A, EP^B) = \begin{cases} 1, & if \sum_{i=1}^{53} MEV_i^A = \sum_{i=1}^{53} MEV_i^B = 0 \\ 1 - \dfrac{\sum_{i=1}^{53} \left| MEV_i^A - MEV_i^B \right|}{\sum_{i=1}^{53} (MEV_i^A + MEV_i^B)}, & otherwise \end{cases} \quad [6.2]$$

Example 6.1. Assume that Genes *A*,*B*,*C*,*D*,*E*,*F* respectively have the following expression profiles across two tissues: $EP^A = [1,1]$, $EP^B = [2,2]$, $EP^C = [3,3]$, $EP^D = [1,2]$, $EP^E = [1,99]$, $EP^F = [1,100]$. The ground-truth similarity relationships that we can intuitively infer include $sim(EP^C, EP^A) < sim(EP^C, EP^B) < 1$, and

$sim(EP^A, EP^D) < sim(EP^E, EP^F) < 1$. Only the results computed by the Bray-Curtis Similarity are completely concordant with these ground-truth relationships (Table 6.2).

Table 6.1: Comparison between metrics in measurement of similarity between gene expression profiles

Similarity metric	Property 1[†, ‡]	Property 2[†]	Property 3[†]
Bray-Curtis	√; [0,1]	√	√
Euclidean	√; (0,1]	√	×
Cosine	√; [0,1]	×	√
Pearson correlation [40]	×; [-1,1]	×	×
Spearman correlation [41]	×; [-1,1]	×	×

[†] √ and × respectively indicate that the similarity metric satisfies and does not satisfy the property.

[‡] The interval in each cell indicates the range in which the result computed by the similarity metric lies.

Table 6.2: Similarity values computed by different metrics

Similarity metric	$sim(EP^C, EP^B)$	$sim(EP^C, EP^A)$	$sim(EP^E, EP^F)$	$sim(EP^A, EP^D)$
Bray-Curtis	0.8	0.5	≈ 0.995	0.8
Euclidean	≈ 0.41	≈ 0.26	0.5	0.5
Cosine	1	1	≈ 0.999999995	≈ 0.949
Pearson correlation	Undefined	Undefined	1	1
Spearman correlation	1	1	1	1

6.2.2 Prediction of Genes with Similar Expression Profiles

The framework for identifying genes that have similar expression profiles to a specific gene is shown in Figure 6.1A and 6.1B. All DHSs in 95 cell types generated by the ENCODE project [19; hg38 assembly] were intersected with known promoters [20], then 94 iPWMs exhibiting primary binding motifs for 82 TFs [3] were used to detect TFBSs in overlapping intervals. When detecting heterotypic TFBS clusters with the IDBC algorithm, a minimum threshold $0.1 * R_{sequence}$ was set for R_i values of TFBSs, in order to remove weak binding sites that were likely to be false positive TFBSs.

The information density-related features derived from each TFBS cluster include: 1) The distance between this cluster and the transcription start site (TSS); 2) The length of this cluster; 3) The information content of this cluster (i.e. the sum of R_i values of all TFBSs in this cluster); 4) The number of binding sites of each TF within this cluster; 5) The number of strong binding sites ($R_i > R_{sequence}$) of each TF within this cluster; 6) The sum of R_i values of binding sites of each TF within this cluster; 7) The sum of R_i values of strong binding sites ($R_i > R_{sequence}$) of each TF within this cluster.

For a gene instance, each of Features 1-3 is defined as a vector whose size equals the number of clusters in the promoter; thus, the entire vector could be input into a classifier. If two instances contained different numbers of clusters, the maximum number of clusters among all instances was determined, and null clusters were added at the 5' end of promoters with fewer clusters, enabling all instances to have the same cluster count. Machine learning classifiers with default parameters in MATLAB were used to generate ROC (Receiver Operating Characteristic) curves.

6.2.3 Prediction of Differentially Expressed Direct TF Targets

6.2.3.1 Using Gene Expression in the CRISPR-based Perturbations

Dixit et al. performed CRISPR-based perturbation experiments using multiple guide RNAs for each of ten TFs in K562 cells, resulting in a regulatory matrix of coefficients that indicate the effect of each guide RNA on each of 22,046 genes [15]. The coefficient

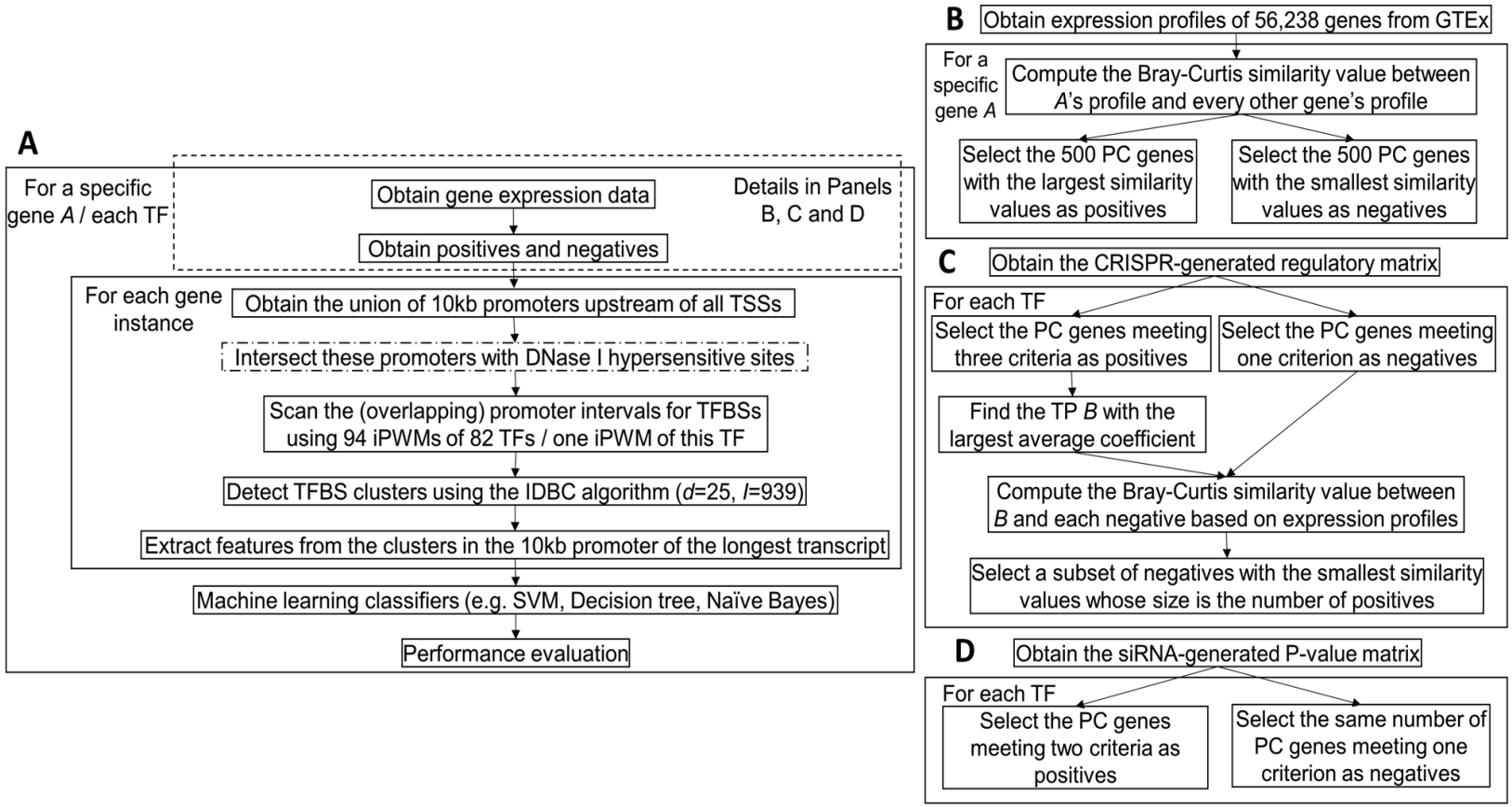

Figure 6.1: The general framework for predicting genes with similar tissue-wide expression profiles and TF targets. A) An overview of the machine learning framework. The steps enclosed in the dashed rectangle and for forming training and test sets vary across prediction of genes with similar expression profiles and DE direct TF targets. The step with a dash-dotted border that intersects promoters with DHSs is a variant of the primary approach that provided more accurate results. In the IDBC algorithm (Appendix E.1), the parameter *I* is the minimum threshold on the total information contents of TFBS clusters. In prediction of genes with similar

expression profiles, the minimum value was 939, which was the sum of mean information contents ($\boldsymbol{R_{sequence}}$ values) of all 94 iPWMs; in prediction of direct targets, this value was the $\boldsymbol{R_{sequence}}$ value of the single iPWM used to detect TFBSs in each promoter. The parameter *d* is the radius of initial clusters in base pairs, whose value, 25, was determined empirically. Eight types of three different classifiers were evaluated with statistics (accuracy, sensitivity and specificity) to measure the classifier performance (Appendix E.1). B) Obtaining of the positives and negatives for identifying genes with similar expression profiles to a given gene (Appendix E.2). C) Obtaining of the positives and negatives for predicting target genes of seven TFs using the CRISPR-generated perturbation data in K562 cells (Appendix E.3). D) Formation of the positives and negatives for predicting target genes of 11 TFs using the siRNA-generated knockdown data in GM19238 cells (Appendix E.4).

of a guide RNA on a gene is defined as the $\log_{10}$(fold change in gene expression level) [15]. Among these ten TFs, we have previously derived iPWMs exhibiting primary binding motifs for seven (EGR1, ELF1, ELK1, ETS1, GABPA, IRF1, YY1) [3]. Therefore, the framework for predicting TF targets in the K562 cell line (Figure 6.1A and 6.1C) was applied to these TFs. The criteria for defining a positive (i.e. a target gene), of a TF was:

1) The fold change in the expression level of this PC gene for each guide RNA of the TF was > (or <) 1, consistent with the possibility that the gene was regulated by the TF, and

2) The average fold change in the expression level of this PC gene for all guide RNAs of the TF was > threshold ε (or < $1/\varepsilon$), and

3) The promoter interval (10 kb) upstream of a TSS of this PC gene overlaps a ChIP-seq peak of the TF in the K562 cell line.

If the coefficients of all guide RNAs of the TF for a PC gene are zero, the gene was defined as a negative. As the threshold ε increases, the number of positives strictly decreases; as ε decreases, we have increasingly lower confidence in the fact that the positives were indeed differentially expressed because of the TF perturbation. To achieve a balance between sensitivity and specificity, we evaluated three different values (i.e. 1.01, 1.05 and 1.1) for ε. For each TF, all ENCODE ChIP-seq peak datasets from the K562 cell line were merged to determine positives. To make the numbers of negatives and positives equal to avoid imbalanced datasets that significantly compromise the classifier performance [21], the Bray-Curtis function was applied to compute the similarity values in the expression profile between all negatives and the positive with the largest average coefficient, then the negatives with the smallest values were selected (Figure 6.1C).

The DHSs in the K562 cell line were intersected with known promoters. Because TFs may exhibit tissue-specific sequence preferences due to different sets of target genes and binding sites in different tissues [3], the iPWMs of EGR1, ELK1, ELF1, GABPA, IRF1,

YY1 from the K562 cell line were used to most accurately detect binding sites; for ETS1, we used the only available iPWM from the GM12878 cell line [3]. Six features were derived from each homotypic cluster (i.e. Features 3 and 6 converged to the same value, because only binding sites from a single TF were used).

6.2.3.2 Using Gene Expression in the siRNA-based Knockdown

In the GM19238 cell line, 59 TFs were individually knocked down using siRNAs, and significant changes in the expression levels of 8,872 genes were indicated according to their corresponding P-values [13]. In these cases, the P-value of a gene for a TF is the probability of observing the change in the expression level of this gene under the null hypothesis of no differential expression after TF knockdown; thus the larger the change in the expression level, the smaller the P-value and the more likely this gene is differentially expressed. They also indicated whether the promoters of these genes display evidence of binding to TFs by intersecting with ChIP-seq peaks in the GM12838 cell line. Among these 59 TFs, we have previously derived accurate iPWMs exhibiting primary binding motifs for 11 (BATF, JUND, NFE2L1, PAX5, POU2F2, RELA, RXRA, SP1, TCF12, USF1, YY1) [3]. Therefore, the framework for predicting TF targets in the GM19238 cell line (Figure 6.1A and 6.1D) was applied to these 11 TFs.

We defined a positive (i.e. a target gene) for a TF, if the P-value of this gene for the TF was ≤ 0.01, and the promoter interval (10kb) upstream of a TSS of this gene overlapped a ChIP-seq peak of the TF in the GM12878 cell line. A negative for a TF exhibited the following property: a P-value > 0.01 for the TF (Figure 6.1D).

The DHSs in the GM19238 cell line mapped from the hg19 genome assembly were first remapped to the hg38 assembly using liftOver (available at genome.ucsc.edu) prior to intersecting with known promoters [22]. Aside from RELA and NFE2L1, the iPWMs of TFs from the GM12878 cell line were used to detect binding sites. For RELA, the iPWM from the GM19099 cell line was used; for NFE2L1, the only available iPWM was derived from K562 cells and was applied. Although the knockdown was performed in GM19238, GM12878 and GM19099 are also lymphoblastic cell lines, with GM19099 and GM19238 both being derived from Yorubans. For this analysis, the iPWMs derived

in GM12878 and GM19099 were more appropriate than the iPWM from K562, since GM12878 and GM19099 are of the same tissue type and are thus more likely comparable to GM19238 than to K562.

6.2.4 Mutation Analyses on Promoters of Differentially Expressed Direct Targets

To better understand the significance of individual binding sites for information-dense clusters and the regulatory state of direct targets, we evaluated the effects of sequence changes that altered the R_i values of these sites on cluster formation and whether a gene was predicted to be a TF target. Mutations were sequentially introduced into the strongest binding sites in TFBS clusters of the EGR1 target gene, *MCM7*, to determine the threshold for cluster formation after disappearing clusters disabled induction of *MCM7* expression. For one target gene of each TF from the CRISPR-generated perturbation data, effects of naturally occurring TFBS variants present in dbSNP (https://www.ncbi.nlm.nih.gov/projects/SNP/) [23] were also evaluated to explore aspects of TFBS organization that enabled both clusters and promoter activity to be resilient to binding site mutations. This was done by analyzing whether the occurrence of individual or multiple single nucleotide polymorphisms (SNPs) lead to the loss of binding sites and the clusters that contain them, and result in changes in the predictions of these targets.

6.3 Results

6.3.1 Similarity between Gene Expression Profiles

To confirm that the Bray-Curtis Similarity can indeed effectively measure how akin the expression profiles of two genes are to each other, it was applied to compute the similarity values between the expression profiles of the glucocorticoid receptor (*GR* or *NR3C1*) gene and all other 56,237 genes. *NR3C1* is an extensively characterized TF with many known direct target genes [24]. As a constitutively expressed TF activated by glucocorticoid ligands, it can mediate the up-regulation of anti-inflammatory genes by binding of homodimers to glucocorticoid response elements and down-regulation of proinflammatory genes by complexing with other activating TFs (e.g. NFKB and AP1) and eliminating their ability to bind targets [24]. NR3C1 can bind its own promoter

forming an auto-regulatory loop, which also contains functional binding sites of 11 other TFs (e.g. SP1, YY1, IRF1, NFKB) whose iPWMs have been developed and/or mutual interactions have been described in Lu et al. [3,24]. However, the expression profile of *NR3C1* integrates all different splicing and translational isoforms (e.g. *GRα-A* to *GRα-D*, *GRβ*, *GRγ*, *GRδ*), whereas these isoforms have tissue-specific expression patterns (e.g. levels of the *GRα-C* isoforms are significantly higher in the pancreas and colon, whereas levels of *GRα-D* are highest in spleen and lungs) [24]. *SLC25A32* and *TANK* have the greatest similarity values to *NR3C1* (0.880 and 0.877 respectively), which is evident intuitively based on their overall similar expression patterns across the 53 tissues (Figure 6.2).

6.3.2 Prediction of Genes with Similar Expression Profiles

In prediction of genes with similar expression profiles to *NR3C1*, we generated ROC curves to compare the performance of different classifiers (Naïve Bayes, Decision Tree, Random Forest and Support vector machines (SVM) with four different kernels), under two scenarios depending on whether promoter sequences were first intersected with DHSs (Figure 6.3). Decision Tree (DT) exhibited the largest AUC under both scenarios, and was one of two most stable classifiers (i.e. $\Delta AUC < 0.01$), with the other being the SVM with RBF kernel. Inclusion of DHS information significantly improved other classifiers' AUC except for Naïve Bayes, and generally all TFBSs in a DHS formed a binding site cluster.

6.3.3 Prediction of Differentially Expressed Direct TF Targets

The best-performing DT classifier in distinguishing genes with *NR3C1*-like expression profiles from others was used to predict TF targets respectively based on the CRISPR- [15] and siRNA-generated [13] perturbation data.

After eliminating TFBSs in inaccessible promoter intervals, i.e. those excluded from tissue-specific DHSs, the DT classifier predicted TF targets with greater sensitivity and specificity (Tables 6.3 and 6.4). Specifically, predictions based on CRISPR-generated knockdown data for TFs: EGR1, ELK1, ELF1, ETS1, GABPA, and IRF1 were more accurate than for YY1, which itself represses or activates a wide range of promoters

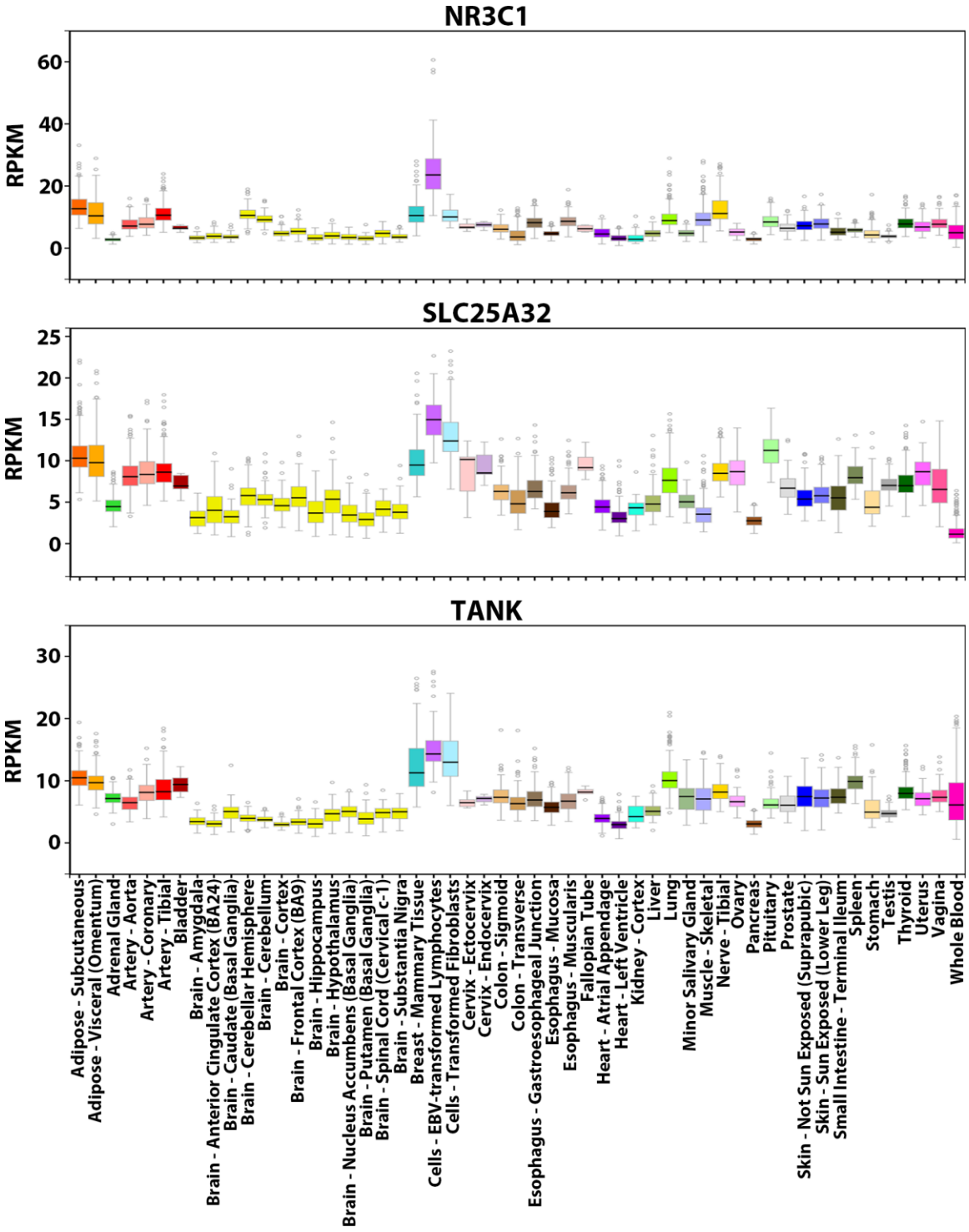

Figure 6.2: Expression profiles of NR3C1, SLC25A32 and TANK. Visualization of the expression values (in RPKM) of these genes across 53 tissues from GTEx. For each gene, the colored rectangle belonging to each tissue indicates the valid RPKM of all samples in the tissue, the black horizontal bar in the rectangle indicates the median

RPKM, the hollow circles indicate the RPKM of the samples considered as outliers, and the grey vertical bar indicates the sampling error. By comparing the pictures, the overall expression patterns of the three genes across the 53 tissues resemble each other (e.g. all three genes exhibit the highest expression levels in lymphocytes and the lowest levels in brain tissues).

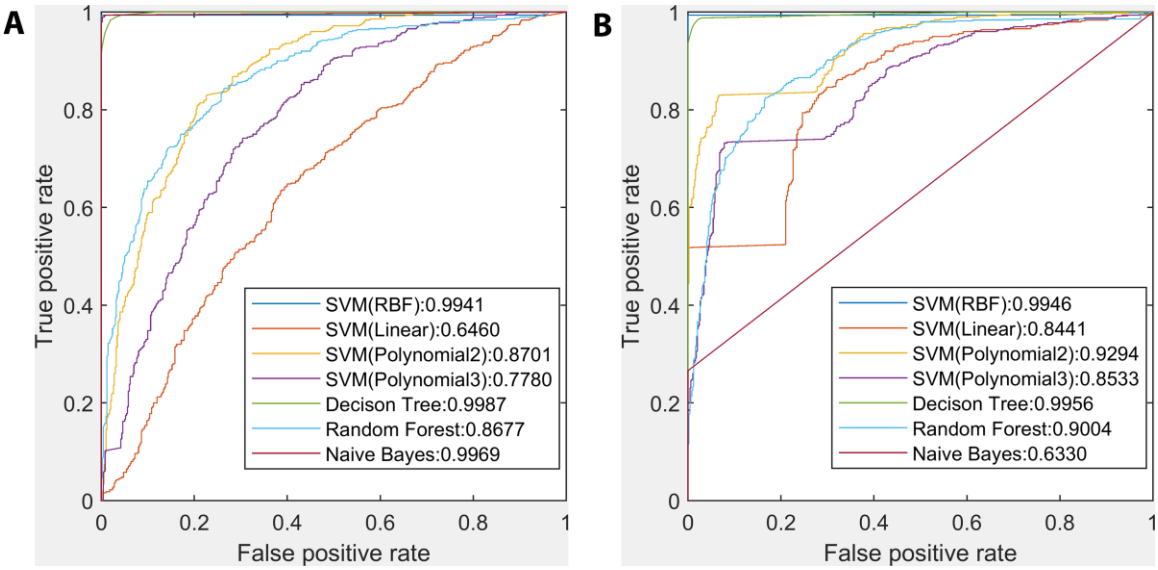

Figure 6.3: Comparison between the performance of different classifiers in prediction of genes with similar expression profiles to NR3C1. (A) ROC curves and AUC of seven classifiers without intersecting promoters with DHSs. (B) ROC curves and AUC of seven classifiers after intersecting promoters with DHSs. The Decision tree classifier exhibited the largest AUC under both scenarios, and inclusion of DHS information significantly improved other classifiers' AUC except for Naïve Bayes.

Table 6.3: The Decision tree classifier performance for predicting TF targets using the CRISPR-generated knockdown data

TF†	Excluding DHS information†			Including DHS information†		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
EGR1	0.58	0.62	0.60	0.78	0.81	0.80
ELF1	0.59	0.65	0.62	0.83	0.87	0.85
ELK1	0.59	0.59	0.59	0.8	0.81	0.81
ETS1	0.59	0.6	0.59	0.81	0.81	0.81
GABPA	0.55	0.57	0.56	0.72	0.75	0.74
IRF1	0.54	0.55	0.54	0.76	0.64	0.70
YY1	0.5	0.51	0.51	0.45	0.69	0.57

† The average performance of 10 rounds of 10-fold cross validation when setting ε to 1.05 is indicated. The CRISPR-generated knockdown data were obtained from Dixit et al [15].

by binding to sites overlapping the TSS (Table 6.3). Accordingly, the perturbation data indicated that YY1 has ~4-22 times more PC targets in the K562 cell line than the other TFs (ε = 1.05), and its binding has a more significant impact on the expression levels of target genes (for YY1, the ratio of the target counts at ε = 1.1 vs ε = 1.01 was 0.334, which significantly exceeded those of the other TFs (0.017-0.082); Appendix E.3). This is concordant with our previous finding that YY1 extensively interacts with 11 cofactors (e.g. DNA-binding IRF9 and TEAD2; non-DNA-binding DDX20 and PYGO2) in K562 cells, consistent with a central role in specifying erythroid-specific lineage development [3].

Despite a high accuracy of target recognition, sensitivity did not exceed specificity except for IRF1 (Table 6.3), due to a relatively large number of false negative genes. Promoters of most TF targets contain accessible, functional binding sites that significantly change gene expression levels upon binding. By contrast, promoters of non-targets contain either no accessible binding sites at all, or accessible, but non-functional sites. The fact that

Table 6.4: The Decision tree classifier performance for predicting TF targets using the siRNA-generated knockdown data

TF†	Excluding DHS information†			Including DHS information†		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
BATF	0.625	0.646	0.604	0.706	0.649	0.763
JUND	0.625	0.646	0.604	0.682	0.682	0.682
NFE2L1	0.633	0.533	0.733	0.75	0.667	0.833
PAX5	0.575	0.614	0.537	0.627	0.563	0.691
POU2F2	0.725	0.818	0.633	0.651	0.796	0.505
RELA	0.591	0.619	0.563	0.690	0.611	0.770
RXRA	0.731	0.813	0.648	0.663	0.793	0.533
SP1	0.561	0.571	0.551	0.579	0.539	0.620
TCF12	0.564	0.638	0.491	0.684	0.597	0.770
USF1	0.737	0.753	0.721	0.723	0.71	0.735
YY1	0.611	0.456	0.765	0.601	0.396	0.807

† The average performance of 10 rounds of 10-fold cross validation is indicated. The siRNA-generated knockdown data were obtained from Cusanovich et al [13].

these false negatives were erroneously predicted to non-targets was attributable to the indistinguishability between functional binding sites in their promoters and non-functional ones in non-targets in the classifier. *In vivo* co-regulation mediated by interacting cofactors, which was excluded by the classifier, assisted in distinguishing these non-functional sites that do not significantly affect gene expression [3,13].

As the threshold ε increased, the accuracy of the classifier for all the TFs monotonically increased as expected (Figure 6.4). For a gene to be defined as a DE target of a TF, the average fold change in its expression level for all guide RNAs that downregulated the TF were required to reach the minimum threshold ε. Upon TF knockdown, ε is inversely

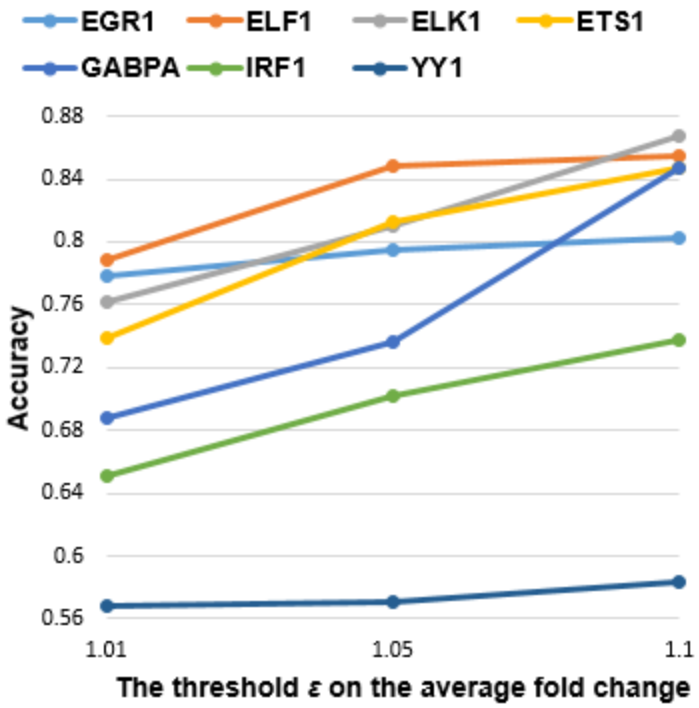

Figure 6.4: Accuracy of the Decision tree classifier when using three different values for ε. Each accuracy value was averaged from 10 rounds of 10-fold cross validation, when the minimum threshold ε on the average fold change in gene expression levels under all guide RNAs of the TF took three different values 1.01, 1.05 and 1.1. As $\boldsymbol{\varepsilon}$ increased, accuracy for all seven TFs monotonically increased.

correlated with the number of target genes, but positively correlated with fold changes in their corresponding expression levels. In general, more significantly DE genes have been associated with a higher number of TFBSs in their promoters [13]. Thus, at greater ε, there are larger differences in the values of machine learning features derived from TFBS clusters between targets and non-targets.

With the siRNA-generated knockdown data, the performance of the DT classifier was compared to the approach inferring DE targets by correlating TF binding with gene expression levels across ten cell types [14]. In this correlation-based approach, three measures (i.e. the absolute Pearson correlation coefficient (PC), the absolute Spearman correlation coefficient (SC), and the absolute combined angle ratio statistic (CARS)), whose performance was evaluated with precision-recall curves, were alternatively used to

compute a correlation score between the number of ChIP-seq peaks overlapping the promoter and gene expression values. Genes predicted to be DE targets had scores above the threshold resulting in a 1.5-fold increase compared to the background precision. For example, in the case of YY1, which was the only TF analyzed by both approaches, the performance of the DT classifier was 0.66 (precision) and 0.456 (recall) (Table 6.4). This classifier outperformed all three correlation measures (PC: 0.467 and 0.003; SC: 0.467 and 0.006; CARS: 0.467 and 0.003), even though the correlation approach used a less stringent P-value threshold (0.05) for defining differential expression of likely non-direct targets, and intersected ChIP-seq peaks over shorter 5kb promoter intervals upstream of the TSS.

6.3.4 Intersection of Genes with Similar Expression Profiles and Direct Targets

To determine how many TF targets have similar tissue-wide expression profiles, we intersected the set of targets with the set of 500 PC genes with the most similar expression profiles for each TF (Table 6.5, Appendix E.5). The TFs PAX5 and POU2F2 are primarily expressed in B cells, and their respective targets *IL21R* and *CD86* are also B cell-specific, which accounts for the high similarity in the expression profile between them. There are respectively 21 and 7 nuclear mitochondrial genes (e.g. *MRPL9* and *MRPS10*, which are subunits of mitochondrial ribosomes) in the intersections for YY1 in the K562 and GM19238 cell lines [25]. Previous studies reported that YY1 upregulates a large number of mitochondrial genes by complexing with PGC-1α in C2C12 cells [26], and genes involved in the mitochondrial respiratory chain in K562 cells [15], which is consistent with the idea that YY1 may broadly regulate mitochondrial function (within all 53 tissues in addition to the erythrocyte, lymphocyte and skeletal muscle cell lines).

Between 0.4%-25% of genes with similar expression profiles to the TFs are actually their targets (Table 6.5); the majority are non-targets whose promoters contain non-functional binding sites that are distinguished from targets by their lack of coregulation by corresponding cofactors. For YY1 and EGR1, we validated this hypothesis by contrasting the flanking cofactor binding site distributions and strengths in the promoters of the most

similarly expressed target genes (YY1: *MRPL9*, *BAZ1B*; EGR1: *CANX*, *NPM1*) and non-target genes (YY1: *ADNP*, *RNF25*; EGR1: *AC142293.3*, *AP000705.7*). Strong and intermediate recognition sites for TFs: SP1, KLF1, CEBPB formed heterotypic clusters with adjacent YY1 sites; as well TFBSs of SP1, KLF1, and NFY were frequently present adjacent to EGR1 binding sites. These patterns contrasted with the enrichment of CTCF and ETS binding sites in gene promoters of YY1 and EGR1 non-targets (Appendix E.6). Previous studies have reported that KLF1 is essential for terminal erythroid differentiation and maturation [27], direct physical interactions between YY1 and the constitutive activator SP1 synergistically induce transcription [28], the activating CEBPB promotes differentiation and suppresses proliferation of K562 cells by binding the promoter of the G-CSFR gene encoding a hematopoietin receptor [29], EGR1 and SP1 synergistically cooperate at adjacent non-overlapping sites on EGR1 promoter but compete binding at overlapping sites [30]; whereas CTCF functions as an insulator blocking the effects of *cis*-acting elements and preventing gene activation [31,32], and ETV6, a member of the ETS family, is a transcriptional repressor required for bone marrow hematopoiesis and associated with leukemia development [33].

6.3.5 Mutation Analyses on Promoters of Direct Targets

Because the promoters of most direct targets contain multiple binding site clusters, we anticipate that this enables these genes' expression to be naturally robust against binding site mutations; in other words, the other clusters can compensate for the loss of a cluster destroyed by mutations in binding sites, so that the mutated promoters are still capable of effectively inducing gene transcription upon TF binding. First, we validated this hypothesis by examining whether introducing artificial variants into binding sites in the promoter of the target gene *MCM7* in the test set of EGR1 changes the classifier output (Figure 6.5). Specifically, in the K562 cell line, *MCM7* is upregulated by EGR1. Knockdown of *MCM7* has an anti-proliferative and pro-apoptotic effect on K562 cells [34] and the loss of EGR1 increases leukemia initiating cells [35], which suggests that EGR1 may act as a tumor suppressor in K562 cells through the *MCM7* pathway.

First, the strongest binding site (at position chr7:100103347 [hg38], - strand, R_i = 12.0 bits) in the promoter was eliminated by a G->A mutation, resulting in the disappearance

Table 6.5: Intersection of TF targets and 500 protein-coding genes with the most similar expression profiles

TF	Cell line	Number of targets	Size of intersection	Targets among the most similar 10 genes[§]
EGR1	K562	169	12	None
ELF1		78	5	None
ELK1		112	4	GNL1(8^{th})
ETS1		267	15	None
GABPA		513	25	TAF1(1^{st})
IRF1		457	10	None
YY1		1752	127	MRPL9(2^{nd}), BAZ1B(6^{th}), ENY2(7^{th}), NUB1(8^{th}), USP1(9^{th}), HNRNPR(10^{th})
	GM19238	1066	61	MED4(1^{st}), SURF6(3^{rd}), BAZ1B(6^{th})
BATF		193	4	MB21D1(4^{th}), C16orf87(9^{th})
JUND		44	2	None
NFE2L1		60	3	None
RELA		252	22	HMG20B(9^{th})
RXRA		183	7	None
SP1		1630	96	ACLY(1^{st}), SEC22B(7^{th}), GPX1P1(10^{th})

TCF12	669	19	None
USF1	309	20	None
PAX5	938	76	IL21R(9th)
POU2F2	550	21	CD86(3rd)

§ The rank of each target in the list of similar genes in the descending order of Bray-Curtis similarity values is shown in the brackets immediately following the target.

of Cluster 1, which consists of two sites (the other site at chr7:100103339, -, 4.3 bits). EGR1 was still predicted to compensate for this mutation, due to the presence of the other two clusters comprising weaker binding sites of intermediate strength (chr7:100102252, +, 7.0 bits; chr7:100102244, +, 1.3 bits; chr7:100101980, +, 8.9 bits; chr7:100101977, +, 2.2 bits; chr7:100101984, +, 1.9 bits), enabling the promoter to maintain capability of inducing *MCM7* expression (Figure 6.5). These adjacent clustered sites, which may not be strong enough to bind TFs and individually activate transcription, can stabilize each other's binding [2]. The weaker sites flanking a strong binding site in a cluster can direct the TF molecule to the strong site and extend the period of the molecule physically associating with the strong site, which is termed, the funnel effect [2]. Further, Clusters 2 and Cluster 3 were respectively removed by G->T and C->G mutations abolishing the strongest site in either cluster, which altered the prediction, that is, EGR1 lost the capability to induce *MCM7* transcription (Figure 6.5). The remaining four sparse weak sites do not form a cluster and cannot completely supplant the disrupted strong sites.

Further, we examined the impacts of known natural SNPs on binding site strengths, clusters and the regulatory state of the promoter for a direct target of each of the seven TFs from the CRISPR-generated perturbation data (Table 6.6). Often a single SNP (e.g. rs996639427 of EGR1) can affect the strengths of multiple binding sites (Table 6.6). Apart from SNPs that are predicted to abolish binding (Figure 6.5), leaky variants that merely weaken TF binding are common (Table 6.6). Binding stabilization between adjacent sites and the funnel effect enable the CRMs comprised of information-dense clusters to be robust to mutations in individual binding sites. In this way, neither mutations that abolish TFBSs nor leaky SNPs in flanking weak sites can destroy functional clusters (e.g. rs1030185383 and rs5874306 of IRF1), whereas SNPs with large reductions in R_i values of central strong sites are more likely to abolish clusters (e.g. rs865922947, rs946037930, rs917218063 and rs928017336 of YY1) (Table 6.5). More generally, the presence of multiple clusters enables promoters to be effectively resilient to the effects binding site mutations; only the complete abolishment of all clusters resulting from the simultaneous occurrence of multiple SNPs can transform the promoter to be unresponsive to TF binding to residual weak sites (e.g. rs997328042, rs1020720126 and rs185306857 of GABPA) (Table 6.6). Furthermore, a relatively small number of SNPs

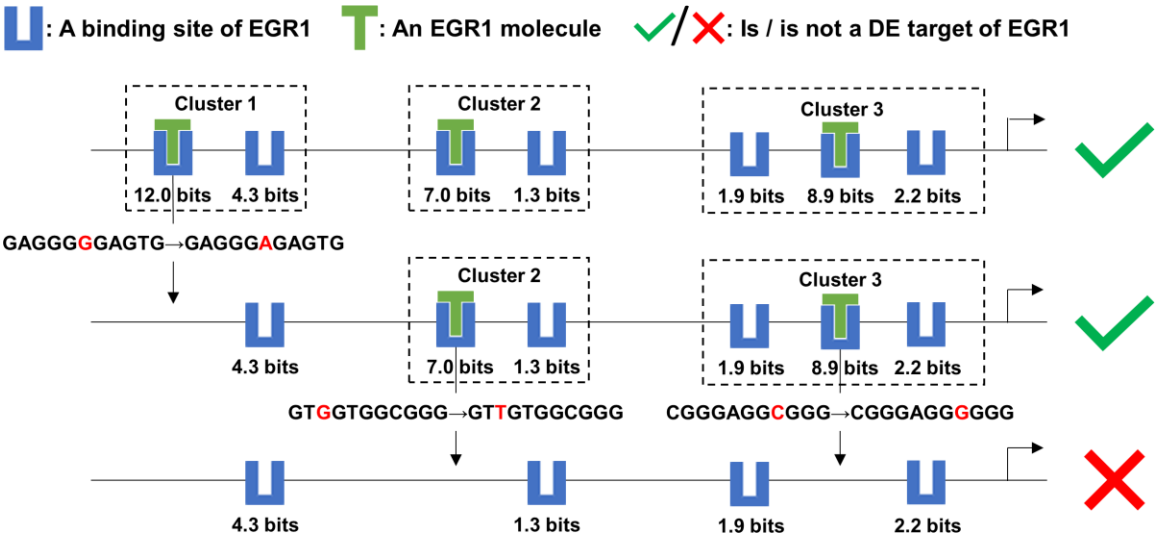

Figure 6.5: Mutation analyses on the target MCM7 in the test set of EGR1. This figure depicts the effect of a mutation in each EGR1 binding site cluster of the *MCM7* promoter on the expression level of *MCM7*, which is a target of the TF EGR1. The strongest binding site in each cluster were abolished by a single nucleotide variant. Upon loss of all three clusters, only weak binding sites remained and EGR1 was predicted to no longer be able to effectively regulate *MCM7* expression. Multiple clusters in the promoters of TF targets confers robustness against mutations within individual binding sites that define these clusters.

that strengthen TF binding and eventually amplify the regulatory effect of the TF on the gene expression level are also present (e.g. rs887888062 of EGR1 and rs751263172 of ELF1) (Table 6.6), suggesting that, in addition to deleterious mutations, benign variants may also be found in promoters, consistent with the expectations of neutral theory [36].

6.4 Discussion

In this study, the Bray-Curtis Similarity function was initially shown (for the *NR3C1* gene) to measure the relatedness of overall expression patterns between genes across a diverse set of tissues. The resulting machine learning framework distinguished Bray-Curtis function-defined similar from dissimilar genes based on the distribution, strengths and compositions of TFBS clusters in accessible promoters, which can substantiallyaccount for the corresponding gene expression patterns. Using knockdown

Table 6.6: Mutation analyses on promoters of direct targets

TF	Target	Normal cluster	Normal allele[§]	SNP ID[§]	Variant allele[§]	Variant cluster[‡]	Classifier output			
							Variant[†]	Wild-type		
EGR1 ($\boldsymbol{R}_{\boldsymbol{sequence}}$ = 12.2899 bits)	EID2B	Cluster 1 of 2	***G***A***G***GGGGC***AT***C (chr19:39540286, -, 7.22 bits)	rs538610162 (chr19:39540296 C>G)	***C***AGGGGGCATC (chr19:39540286, -, 4.84 bits)	Abolished	√			
				rs759233998 (chr19:39540294 C>T)	GA***A***GGGGCATC (chr19:39540286, -, 0.06 bit)	Abolished	√			
				rs974735901 (chr19:39540288 T>A)	GAGGGGGC***T***TC (chr19:39540286, -, 6.90 bits)	Cluster 1 of 2	√	×	√	
				rs978230260 (chr19:39540287 A>T)	GAGGGGGCA***A***C (chr19:39540286, -, 5.31 bits)	Abolished	√			
		Cluster 2 of 2	***G***CGTGCGT***G***GG (chr19:39540162, +, 1.59 bits)	rs764734511 (chr19:39540162 G>A)	***A***CGTGCGTGGG (chr19:39540162, +, -0.72 bit)	Cluster 2 of 2	√			
				(chr19:39540162	***C***CGTGCGTGGG	Cluster 2	√			

				G>C)	(chr19:39540162, +, -0.79 bit)	of 2		
				rs996639427 (chr19:39540170 G>C)	GCGTGCGT***C***GG (chr19:39540162, +, -5.21 bits)	Abolished	√	
					GCGT***C***GGCGCT (chr19:39540165, +, -0.85 bit)			
			GCGT***G***GGC***G***C***T*** (chr19:39540166, +, 9.72 bits)	rs1027751538 (chr19:39540174 G>A)	GCGTGGGC***A***CT (chr19:39540166, +, 5.16 bits)	Abolished	√	
				<u>rs887888062 (chr19:39540176 T>A)</u>	<u>GCGTGGGCGC***A*** (chr19:39540166, +, 10.94 bits)</u>	Cluster 2 of 2	√	
ELF1 (***R***$_{sequence}$ = 11.2057 bits)	HIST1H 4H	Cluster 1 of 2	GC***GGA***AG***CG***TG (chr6:26286540, +, 9.92 bits)	<u>rs760968937 (chr6:26286547 C>T)</u>	<u>GCGGAAG***T***GTG (chr6:26286540, +, 10.71 bits)</u>	Cluster 1 of 2	√	√	√
				rs760968937 (chr6:26286547 C>A)	GCGGAAG***A***GTG (chr6:26286540, +, 8.84 bits)	Cluster 1 of 2	√	×	

		rs1000196206 (chr6:26286542 G>C)	GC***C***GAAGCGTG (chr6:26286540, +, -6.26 bits)	Abolished	√
		rs144759258 (chr6:26286543 G>A)	GCG***A***AAGCGTG (chr6:26286540, +, -3.60 bits)	Abolished	√
		rs966435996 (chr6:26286544 A>G)	GCGG***G***AGCGTG (chr6:26286540, +, 5.28 bits)	Abolished	√
		rs950986427 (chr6:26286548 G>A)	GCGGAAGC***A***TG (chr6:26286540, +, 8.28 bits)	Cluster 1 of 2	√
		rs373649904 (chr6:26286483 G>A)	***T***AGGAGATGCG (chr6:26286473, -, 0.61 bit)	Abolished	√
Cluster 2 of 2	***C***AG***GAG***ATGC***G*** (chr6:26286473, -, 6.98 bits)	rs926919149 (chr6:26286480 C>T)	CAG***A***AGATGCG (chr6:26286473, -, -6.53 bits)	Abolished	√
		rs751263172 (chr6:26286479	CAGG***C***GATGCG (chr6:26286473,	Abolished	√

				T>G)	-, 1.24 bits)			
				rs369076253 (chr6:26286473 C>G)	CAGGAGATGC***C*** (chr6:26286473, -, 6.92 bits)	Cluster 2 of 2	√	
				rs751263172 (chr6:104447431 4C>T)	CAGGA***A***ATGCG (chr6:26286473, -, 11.43 bits)	Cluster 2 of 2	√	√
ELK1 ($\boldsymbol{R_{sequence}}$ = 11.9041 bits)	G0S2	Cluster 1 of 2	C***AG***GGAAG***A***CC (chr1:209667959, -, 1.92 bits)	rs146048477 (chr1:209667961 T>A)	CAGGGAAG***T***CC (chr1:209667959, -, 2.24 bits)	Cluster 1 of 2	√	√
				rs887606802 (chr1:209667968 T>C)	C***G***GGGAAGACC (chr1:209667959, -, -3.35 bits)	Cluster 1 of 2	√	×
				rs1021034916 (chr1:209667967 C>T)	CA***A***GGAAGACC (chr1:209667959, -, -3.57 bits)	Cluster 1 of 2	√	
			GAGGA***A***ATGAG (chr1:209667969, +, 8.14 bits)	rs941962117 (chr1:209667974 A>G)	GAGGA***G***ATGAG (chr1:209667969, +, 4.11 bits)	Abolished	√	
		Cluster 2 of	***C***TGGAAGA***GC***A	rs896117033	CTGGAAGAG***T***A	Cluster 2 of	√	

		2	(chr1:209673544, -, 5.91 bits)	(chr1:209673545 G>A)	(chr1:209673544, -, 3.95 bits)	2			
				rs971962577 (chr1:209673546 C>T)	CTGGAAGA***A***CA (chr1:209673544, -, 3.49 bits)	Cluster 2 of 2	√		
			CCA***G***AAGTCA***A*** (chr1:209673551, +, 7.44 bits)	rs1011969709 (chr1:209673554 G>C)	***G***TGGAAGAGCA (chr1:209673544, -, 3.92 bits)	Abolished	√		
					CCA***C***AAGTCAA (chr1:209673551, +, -5.50 bits)				
				<u>rs1023312090 (chr1:209673561 A>G)</u>	<u>CCAGAAGTCA***G*** (chr1:209673551, +, 8.40 bits)</u>	Cluster 2 of 2	√	√	
ETS1 (***R*** $_{sequence}$ = 10.0788 bits)	TTC19	Cluster 1 of 1	GCA***G***GGAA***A***GG (chr17:16022293, +, 7.92 bits)	rs1022234223 (chr17:16022296 G>C)	GCA***C***GGAAAGG (chr17:16022293, +, -4.98 bits)	Abolished	×	×	√
				<u>rs968299415 (chr17:16022301 A>T)</u>	<u>GCAGGGAA***T***GG (chr17:16022293, +, 10.01 bits)</u>	Cluster 1 of 1	√	√	

GABPA ($\boldsymbol{R_{sequence}}$ = 10.8567 bits)	PLEKHB2	Cluster 1 of 1	A***C***A***G***GAAAGGG (chr2:131112770, +, 10.36 bits)	rs997328042 (chr2:131112771 C>T)	A***T***AGGAAAGGG (chr2:131112770, +, -3.68 bits)	Abolished	×	
				rs1020720126 (chr2:131112773 G>C)	ACA***C***GAAAGGG (chr2:131112770, +, -4.16 bits)	Abolished	×	×
				rs185306857 (chr2:131112761 C>A)	T***A***TGGAAACTA (chr2:131112760, +, -2.86 bits)	Cluster 1 of 1	√	√
			T***CT***GGAAAC***T***A (chr2:131112760, +, 1.53 bits)	rs772728699 (chr2:131112762 T>A)	TC***A***GGAAACTA (chr2:131112760, +, 5.23 bits)	Cluster 1 of 1	√	
				rs965753671 (chr2:131112769 T>C)	TCTGGAAAC***C***A (chr2:131112760, +, 2.13 bits)	Cluster 1 of 1	√	
IRF1 ($\boldsymbol{R_{sequence}}$ = 13.5544 bits)	SMIM13	Cluster 1 of 1	***GA***G***A***A***T***GAA***A***G C***A*** (chr6:11093663, +, 12.56 bits)	rs950528541 (chr6:11093663 G>C)	***C***AGAATGAAAG CA (chr6:11093663, +, 8.97 bits)	Cluster 1 of 1	√	×	√

	rs886259573 (chr6:11093664 A>G)	G***G***GAATGAAAG CA (chr6:11093663, +, 9.65 bits)	Cluster 1 of 1	√
	rs982931728 (chr6:11093666 A>G)	GAG***G***ATGAAAG CA (chr6:11093663, +, 8.09 bits)	Cluster 1 of 1	√
	rs1020218811 (chr6:11093668 T>G)	GAGAA***G***GAAAG CA (chr6:11093663, +, 9.36 bits)	Cluster 1 of 1	√
	rs570723026 (chr6:11093672 A>G)	GAGAATGAA***G***G CA (chr6:11093663, +, 8.01 bits)	Cluster 1 of 1	√
	rs1004825794 (chr6:11093675 A>C) (chr6:11093675	GAGAATGAAAG C***C*** (chr6:11093663, +, 10.47 bits)	Cluster 1 of 1	√

				A>T)	GAGAATGAAAGC***A*** (chr6:11093663, +, 10.42 bits)	Cluster 1 of 1	√		
			AA***G***ACCAA***AG***GCA (chr6:11093641, +, 2.43 bits)	rs1030185383 (chr6:11093649 A>C)	AAGACCAA***C***GGCA (chr6:11093641, +, -3.39 bits)	Cluster 1 of 1	√		
				rs5874306 (chr6:11093650delG)	AAGACCAAAGCAG (chr6:11093641, +, 0.90 bit)	Cluster 1 of 1	√		
				<u>rs558896490 (chr6:11093643 G>A)</u>	<u>AA***A***ACCAAAGGCA (chr6:11093641, +, 7.06 bits)</u>	Cluster 1 of 1	√	√	
YY1 (***R*** $_{sequence}$ = 12.8554 bits)	CKLF	Cluster 1 of 1	***G***C***GG***C***C***ATCGGC (chr16:66549785, -, 10.06 bits)	rs865922947 (chr16:66549791 G>A)	***C***CGGCCATCGGC (chr16:66549785, -, 6.80 bits)	Cluster 1	√	×	√
				rs946037930	GC***T***GCCATCGGC	Cluster 1	√		

	(chr16:66549794 C>A)	(chr16:66549785, -, 8.02 bits)		
	rs917218063 (chr16:66549793 C>T)	GCG***A***CCATCGGC (chr16:66549785, -, 5.41 bits)	Abolished	×
	rs928017336 (chr16:66549791 G>A)	GCGGC***T***ATCGGC (chr16:66549785, -, -3.62 bits)	Abolished	×
GCCGCCCCCGTC (chr16:66549792, +, 1.34 bits)				

§ All coordinates are based on the hg38 genome assembly. A bold italic letter in a binding site sequence indicates the base where a SNP occurs. The SNPs strengthening binding sites and corresponding variant binding site sequences are underlined.

‡ The impact on whether the occurrence of a single SNP resulted in the disappearance of the cluster containing it is shown.

† After a single SNP occurred or multiple SNPs simultaneously occurred, the classifier produced a new prediction on whether the TF is still capable of significantly affecting gene expression via the variant promoter.

data as the gold standard, the combinatorial use of TF binding profiles and chromatin accessibility was also demonstrated to be predictive of TF targets. A binding site comparison confirmed that coregulatory cofactors are responsible for distinguishing between functional sites in targets and non-functional ones in non-targets. Furthermore, mutation analyses on binding sites of targets demonstrated that the existence of both multiple TFBSs in a cluster and multiple information-dense clusters in a promoter enables both the cluster and the promoter to be resilient to binding site mutations.

The DT classifier improved after intersecting promoters with DHSs in both prediction of genes with similar expression profiles to *NR3C1* and prediction of TF targets (Figure 6.3, Tables 6.3 and 6.4). This intersection eliminated noisy binding sites that are inaccessible to TF proteins in promoters; specifically, it widened discrepancies in feature vectors between positives and negatives. If the 10kb promoter of a gene instance does not overlap DHSs, its feature vector will only consist of 0; the percentages of negatives whose promoters do not overlap DHSs considerably exceeded those of positives (Appendix E.7), which led to an excess of negatives with feature vectors containing only 0 after intersection. This explains why these negatives are not DE targets of the TFs in the K562 and GM19238 cell lines, because their entire promoters are not open to TF molecules; other regulatory regions besides the proximal promoters (e.g. intronic enhancers [37]) still enable the TFs to effectively control the expression of the positives with inaccessible promoters.

The relatively poor performance of the classifier on YY1 (Table 6.3) is attributable to its smaller percentage of negatives with inaccessible promoters and the larger number of functional binding sites in the K562 cell line (Appendix E.7). Additionally, the DT classifier was more predictive of functional TF binding on the CRISPR-generated knockdown data than the siRNA-generated ones (Tables 6.3 and 6.4). This larger discrepancy in feature vectors between positives and negatives from CRISPR-based perturbations is also attributable to the greater differences in the percentages between positives and negatives with inaccessible promoters (Appendix E.7). Among the 22,046 genes whose expression levels were measured in the CRISPR-based perturbations, most of the TNs with inaccessible promoters merely have one transcript and specific functions

(e.g. *VENTXP1* for the TF, EGR1), whereas many such negatives were excluded from the 8,872 genes whose knockdown data were generated by siRNA inactivation.

Our *in-silico* mutation analyses revealed that some deleterious TFBS mutations could be compensated for by other information-dense clusters in a promoter; thus predicting the effects of mutations in individual binding sites would not be sufficient to interpretation of downstream effects. Though compensatory clusters may maintain gene expression, the promoter will provide lower levels of activity than the wild-type promoter could, which is a recipe for achieving natural phenotypic diversity. Few published studies in molecular diagnostics have specifically examined the effects of naturally occurring variants within clustered TFBSs; thus IDBC-based machine learning provided an alternative computational approach to predict deleterious mutations that actually impact (i.e. repress or abolish) transcription of target genes and result in abnormal phenotypes, and to simultaneously minimize false positive calls of TFBS mutations that individually have little or no impact.

Apart from these TFs, the Bray-Curtis Similarity can be directly applied to identify the ground-truth genes with overall similar tissue-wide expression patterns to any other gene whose expression profile is known. Further studies could investigate the biological significance underlying the phenomenon that all these genes share a common expression pattern, including the similarity between other regulatory regions besides proximal promoters in terms of TFBSs and epigenetic markers. This machine learning framework can also be applied to predict direct DE targets for other TFs and in other cell lines, depending on the availability of corresponding knockdown data.

There are a number of limitations of our approach. The Bray-Curtis function seems unable to accurately measure the similarity between the expression profiles of a gene (e.g. *MIR23A*) without any detectable mRNA in any of the 53 tissues analyzed and genes (e.g. the ubiquitously expressed *NR3C1* and stomach-specific *PGA3*) that are expressed in at least one tissue. Intuitively, in terms of expression patterns *PGA3* is more similar to *MIR23A* than *NR3C1*; however, the Bray-Curtis similarity values indicate that both *PGA3* and *NR3C1* bear no similarity to *MIR23A* (i.e. $sim_{Bray-Curtis}(NR3C1, MIR23A) =$

$sim_{Bray-Curtis}(PGA3, MIR23A) = 0$). Another possible limitation in classifier performance in the prediction of genes with similar tissue-wide expression profiles is that only binding sites of 82 TFs were analyzed due to a lack of available iPWMs for other TFs, given that 2000-3000 sequence-specific DNA-binding TFs are estimated to be encoded in the human genome [38]. For example, four TFs (CREB, MYB, NF1, GRF1) were previously reported to bind the promoter of the *NR3C1* gene to activate or repress its expression, however their iPWMs exhibiting known primary motifs could not be successfully derived from ChIP-seq data [3,24]. Regarding the CRISPR-generated knockdown data used here, TPs were inferred to be direct targets by intersecting promoters with their corresponding ChIP-seq peaks, which may not be completely accurate, due to the presence of noise peaks that do not contain true TFBSs [3,39]. In instances where small fold changes in the expression levels of DE targets were evident, these peaks could arise from compromised efficiency of knockdowns as a result of suboptimal guide RNAs or the limitations of perturbing only a single allele of the TF. Finally, the framework developed here only takes into account the 10kb interval proximal to the TSS, and would not therefore capture long range enhancer effects beyond this distance; by contrast, correlation based approaches have successfully incorporated multiple definitions of promoter length [14].

6.5 Conclusions

The Bray-Curtis function is able to effectively quantify the similarity between tissue-wide gene expression profiles. By analysis of promoter information theory-based TF binding profiles that captured the spatial distribution and information contents of TFBS clusters, ChIP-seq and chromatin accessibility data, we described a machine learning framework that distinguished tissue-wide expression profiles of similar vs dissimilar genes and identified direct DE targets of TFs. Functional binding sites in target genes that significantly alter expression levels upon direct binding are at least partially distinguished by TF-cofactor coregulation from non-functional sites in non-targets. Finally, *in-silico* mutation analyses demonstrated that the presence of multiple information-dense clusters in the promoter reduces deleterious mutations that can significantly alter the regulatory state and expression level of the gene as a protective mechanism.

6.6 References

1. Hosseinpour B, Bakhtiarizadeh MR, Khosravi P, Ebrahimie E. Predicting distinct organization of transcription factor binding sites on the promoter regions: a new genome-based approach to expand human embryonic stem cell regulatory network. Gene. 2013;531:212–9.

2. Ezer D, Zabet NR, Adryan B. Homotypic clusters of transcription factor binding sites: A model system for understanding the physical mechanics of gene expression. Comput Struct Biotechnol J. 2014;10:63–9.

3. Lu R, Mucaki EJ, Rogan PK. Discovery and validation of information theory-based transcription factor and cofactor binding site motifs. Nucleic Acids Res. 2017;45:e27.

4. Schneider TD. Information content of individual genetic sequences. J Theor Biol. 1997;189:427–41.

5. Dinakarpandian D, Raheja V, Mehta S, Schuetz EG, Rogan PK. Tandem machine learning for the identification of genes regulated by transcription factors. BMC Bioinformatics. 2005;6:204.

6. Ouyang Z, Zhou Q, Wong WH. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proc Natl Acad Sci U S A. 2009;106:21521–6.

7. Cheng C, Alexander R, Min R, Leng J, Yip KY, Rozowsky J, et al. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome Res. 2012;22:1658–67.

8. Budden DM, Hurley DG, Cursons J, Markham JF, Davis MJ, Crampin EJ. Predicting expression: the complementary power of histone modification and transcription factor binding data. Epigenetics Chromatin. 2014;7:36.

9. Smith AD, Sumazin P, Xuan Z, Zhang MQ. DNA motifs in human and mouse proximal promoters predict tissue-specific expression. Proc Natl Acad Sci U S A. 2006;103:6275–80.

10. McLeay RC, Lesluyes T, Cuellar Partida G, Bailey TL. Genome-wide in silico prediction of gene expression. Bioinforma Oxf Engl. 2012;28:2789–96.

11. Karlić R, Chung H-R, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. Proc Natl Acad Sci U S A. 2010;107:2926–31.

12. Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, et al. Modeling gene expression using chromatin features in various cellular contexts. Genome Biol. 2012;13:R53.

13. Cusanovich DA, Pavlovic B, Pritchard JK, Gilad Y. The functional consequences of variation in transcription factor binding. PLoS Genet. 2014;10:e1004226.

14. Banks CJ, Joshi A, Michoel T. Functional transcription factor target discovery via compendia of binding and expression profiles. Sci Rep. 2016;6:20649.

15. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell. 2016;167:1853-1866.e17.

16. Cui S, Youn E, Lee J, Maas SJ. An improved systematic approach to predicting transcription factor target genes using support vector machine. PloS One. 2014;9:e94519.

17. Bray JR, Curtis JT. An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecol Monogr. 1957;27:325–349.

18. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45:580–5.

19. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

20. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489:75–82.

21. He H, Garcia EA. Learning from Imbalanced Data. IEEE Trans Knowl Data Eng. 2009;21:1263–1284.

22. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12:996–1006.

23. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29:308–11.

24. Vandevyver S, Dejager L, Libert C. Comprehensive overview of the structure and regulation of the glucocorticoid receptor. Endocr Rev. 2014;35:671–93.

25. Calvo SE, Clauser KR, Mootha VK. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. Nucleic Acids Res. 2016;44:D1251-1257.

26. Cunningham JT, Rodgers JT, Arlow DH, Vazquez F, Mootha VK, Puigserver P. mTOR controls mitochondrial oxidative function through a YY1-PGC-1alpha transcriptional complex. Nature. 2007;450:736–40.

27. Tallack MR, Perkins AC. KLF1 directly coordinates almost all aspects of terminal erythroid differentiation. IUBMB Life. 2010;62:886–90.

28. Seto E, Lewis B, Shenk T. Interaction between transcription factors Sp1 and YY1. Nature. 1993;365:462–4.

29. Ferrari-Amorotti G, Mariani SA, Novi C, Cattelani S, Pecorari L, Corradini F, et al. The biological effects of C/EBPalpha in K562 cells depend on the potency of the N-terminal regulatory region, not on specificity of the DNA binding domain. J Biol Chem. 2010;285:30837–50.

30. Huang RP, Fan Y, Ni Z, Mercola D, Adamson ED. Reciprocal modulation between Sp1 and Egr-1. J Cell Biochem. 1997;66:489–99.

31. Bell AC, West AG, Felsenfeld G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell. 1999;98:387–96.

32. Hou C, Zhao H, Tanimoto K, Dean A. CTCF-dependent enhancer-blocking by alternative chromatin loop formation. Proc Natl Acad Sci U S A. 2008;105:20398–403.

33. Wang LC, Swat W, Fujiwara Y, Davidson L, Visvader J, Kuo F, et al. The TEL/ETV6 gene is required specifically for hematopoiesis in the bone marrow. Genes Dev. 1998;12:2392–402.

34. Tian L, Liu J, Xia G-H, Chen B-A. RNAi-mediated knockdown of MCM7 gene on CML cells and its therapeutic potential for leukemia. Med Oncol Northwood Lond Engl. 2017;34:21.

35. Maifrede S, Liebermann D, Hoffman B. Egr-1, a Stress Response Transcription Factor and Myeloid Differentiation Primary Response Gene, Behaves As Tumor Suppressor in CML. Blood. 2014;124:2211.

36. Kimura M. The neutral theory of molecular evolution. Sci Am. 1979;241:98–100, 102, 108 passim.

37. Hural JA, Kwan M, Henkel G, Hock MB, Brown MA. An intron transcriptional enhancer element regulates IL-4 gene locus accessibility in mast cells. J Immunol Baltim Md 1950. 2000;165:3239–49.

38. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009;10:252–63.

39. Kidder BL, Hu G, Zhao K. ChIP-Seq: technical considerations for obtaining high-quality data. Nat Immunol. 2011;12:918–22.

40. Pearson K. Note on Regression and Inheritance in the Case of Two Parents. Proc R Soc Lond. 1895;58:240–2.

41. Spearman C. The Proof and Measurement of Association between Two Things. Am J Psychol. 1904;15:72–101.

Chapter 7

7 Discussion

In this chapter, we will discuss the advances and generalization of our methods developed in Chapters 3, 4, 5 and 6, the implications of our results, and the potential limitations and future studies to overcome them.

7.1 Advances and Generalization of the Methods Developed in this Thesis

7.1.1 The Maskminent Motif Discovery Pipeline and iPWM Validation

The Maskminent motif discovery pipeline we developed in Chapter 3 provides a complete software suite to mine elusive TFBS motifs from ChIP-seq datasets, generate contiguous and bipartite iPWMs, and scan DNA sequences for binding sites using iPWMs.

Compared to Bipad, one advance of the Maskminent pipeline is the ability to recursively mask motifs found in previous iterations to discover additional conserved motifs from the same dataset. These previous motifs are masked by skipping the binding sites with $R_i > 0$ predicted by iPWMs. Bipad can only perform one execution on a ChIP-seq dataset and return the lowest-entropy motif, regardless of whether this motif is recognized by a TF or noise. By contrast, MEME uses a likelihood-based approach to mask the motifs that have been found by the algorithm (1). For each position, MEME computes the probability that it is not contained in an occurrence of any motif found to date, which further affects the next estimates for base frequencies in the motif component in the maximization step (1).

The ability to identify the maximum number of top peaks that can produce the primary or cofactor motif is another advance of the Maskminent pipeline. First, it also enhances the ability of Bipad to reveal primary and cofactor motifs. If Bipad returns a noise motif from all peaks in a dataset, this implies the alignment formed by all binding sites in the dataset has a larger entropy than the noise motif. Since peaks with higher signal values generally

contain stronger binding sites, sites in a subset of top peaks have higher sequence conservation (i.e. more similar to the strongest consensus sequence) and form a lower-entropy alignment. Thus this alignment is more likely to suppress the noise motif and be returned by Maskminent. Second, compared to prior motif discovery pipelines applied only to a small number of top peaks (2–4), the maximum number of top peaks enable the derived iPWM to incorporate as many intermediate and weak binding sites as possible. Thus the iPWM can also accurately quantify the strengths of intermediate and weak binding sites apart from strong sites.

Compared to the previous studies (2–6) deriving contiguous count matrices or PFMs from ChIP-seq datasets, the generation of bipartite iPWMs is an advance. From Equation 1.6, the R_i value of a binding site or a non-site computed from an iPWM is quantitatively related to the amount of heat energy dissipated or absorbed by the association of the TF protein with the site or non-site, enabling the accurate quantification of the binding site strength, in contrast with the log likelihood ratio score computed from a PFM that is not based on Shannon information theory. In addition, bipartite iPWMs are capable of accurately modelling the binding behavior of dimeric TFs by allowing the spacer length to vary and computing a geometry-based spacer penalty.

Using iPWMs to detect experimentally confirmed binding sites and interpret the experimentally determined effects of SNPs on binding site strengths is an advance to the methods in the literature to validate the TFBS motifs derived from ChIP-seq datasets, and ought to be the gold standard. Another method widely used by previous studies (4, 7) is to generate the ROC (receiver operating characteristics) curve and measure the AUC (area under the curve), which is not as convincing. For example, when generating the ROC curve, the negative sequences that supposedly do not contain binding sites were randomly chosen from the genome (7). In fact, these randomly chosen sequences may still contain true binding sites, as demonstrated by the false positive detection rate (1.2E-7 to 0.06) from the null R_i distribution in Chapter 3, which results in an inaccuracy.

In Chapter 3 the Maskminent pipeline was applied to analyze almost all ChIP-seq datasets of human TFs that ENCODE had generated before April 2016. ENCODE has

been conducting more ChIP-seq experiments, resulting in a total of ~2,400 peak datasets to date. The Maskminent pipeline can be directly applied to these newly generated data to derive binding site motifs and iPWMs for more TFs. Furthermore, 2000-3000 sequence-specific DNA-binding TFs are estimated to be encoded in the human genome (8); specifically, the combination of multiple paper and database sources indicated that 2,765 proteins are likely to act as TFs (9). ChIP-seq experiments have not been performed for numerous TFs with known or unknown binding specificities, and even unknown TFs, possibly due to a lack of efficient protein antibodies; for example, 1,211 out of the known or likely 1,639 TFs with high confidence have known binding motifs, with 1,107 determined experimentally and 104 inferred from a closely-related homolog (9). The Maskminent pipeline can also be applied to these future ChIP-seq datasets, with the ultimate objective to determine the sequence preferences for all human TFs.

As the number of available ChIP-seq datasets generated by ENCODE rapidly increases, the scalability and running efficiency of the Maskminent software become particularly important. Maskminent, being a C++ software program, is time efficient despite its output is not transiently generated; for example, on an Intel Xeon processor with a 2.27 GHz clock frequency, it took three hours to find the 22nt-wide optimal multiple alignment from the top 1,000 ~330nt-long peaks of the TF RFX5. As high-performance computing facilities with numerous processors like SHARCNET become increasingly ubiquitous, Maskminent can handle in parallel a large number of ChIP-seq datasets by running independent instances on multiple processors, further reducing the time required to analyze all data; for example, it took approximately seven months to obtain the results from all 765 datasets in Chapter 3 using the Maskminent pipeline on approximately eight processors of SHARCNET, amounting to about five years of CPU time (i.e. a single CPU keeps working for five years without any break). Thus we are able to estimate that it will take ~15 months to finish analyzing these newly generated ~1,635 datasets on the same processors, amounting to ~10 years of CPU time.

7.1.2 The Unified Analytic Framework for Prioritization of Non-coding Variants of Uncertain Significance in Heritable Breast and Ovarian Cancer

One advance of this unified framework is that besides coding variants, it also integrates information theory-based analyses of multiple types of non-coding variants, including TFBS, SRBS, RBBS, and splice site variants. Shannon information theory confers this unifying capability; as long as an iPWM quantifying its base preferences is derived for a regulatory protein of any type (e.g. TFs, splicing regulatory or RNA-binding proteins), we will be able to detect its binding sites and predict the effects of variants within these sites based on changes in the R_i values.

Another advance of this framework is that its predictions on the effects of non-coding variants are robust and accurate, due to the quantitative relationship between Shannon information and the Second Law of Thermodynamics (Equation 1.6). This hypothesis has been proved by the successful detection of true TFBSs in Chapter 3, explanation of experimentally observed effects of SNPs on TFBS strengths in Chapter 3, and interpretation of experimentally determined effects of splice site variants on mRNA splicing (10–13).

In Chapter 4, this unified framework was applied to identify 15,311 unique variants in 7 complete HBOC genes of 102 patients; on a larger scale, Chapter 5 further identified 38,372 variants in 20 HBOC genes of 287 patients. These genes are known to be associated with increased risks of HBOC through four pathways; however, variants in many other genes can also contribute to the onset of HBOC. Furthermore, non-coding genetic mutations are also responsible for the susceptibility to other types of cancer, such as lung cancer (14) and colon cancer (15); however, most of these mutations do not lie within TF-encoding genes based on data from recent CRISPR screens, perhaps because the human TFs mainly serve developmental or tissue-specific functions (9, 16). By contrast, 19.1% of TF-encoding genes were found to harbor mutations associated with at least one disease phenotype, a higher percentage than that observed for all genes (16.2%) (9). This information theory-based framework can incorporate iPWMs of more TFs that

will be derived from the newly generated ChIP-seq datasets and be directly applied to other HBOC genes, cancer types and diseases.

As the cost and time required by whole genome sequencing that sequences the complete genome of an individual rapidly decrease due to the progressive adoption of next-generation sequencing (NGS) technologies, identifying all coding and non-coding variants in a complete genome instead of specific genes becomes feasible, and the scalability of this unified framework becomes important in terms of both hardware consumption and running time. Considering only TFBS variants for simplicity, one can reasonably estimate that ~3 million binding site variants of ~3,000 TFs will be identified in the complete genome of a patient. The hardware resources required for applying the framework to analyze these variants will include a large-capacity disk drive storing the input data (i.e. the ~3 billion bp-long genome, 3 million variants and ~3,000 TF iPWMs) and the analysis results, RAM and processors. Based on the current electronics industry, the amount of available RAM, being several dozen GB in a typical computer system, is more of a bottleneck compared to the disk drive and processors, due to the low hard drive cost and high-performance computing facilities. Such an amount is not sufficient for simultaneously accommodating all the input data and output results, leading to the common space-time trade-off (17) between hardware consumption and running time faced by many computational algorithms (e.g. dynamic programming (18)). One batch-based strategy to approach this issue is to maximize space usage and minimize running time by dividing the input data into as few batches as possible. Each batch contains as much data (TFBS variants in the case of the unified framework) as possible that can be processed by the limited RAM at one time.

7.1.3 The General Machine Learning Framework for Prediction of Gene Expression Profiles and TF Target Genes

Compared to prior studies (19–22) predicting tissue-specific gene expression levels, the definition of the gene expression profile being the expression values across 53 tissues is an advance of our machine learning model. Focusing on one single tissue at one time limits the tissue-wide expandability of the models developed in the prior studies; either a different model needed to be constructed for each tissue (22) or the same model needed

to be rerun for each tissue (19–21). By contrast, this definition of the gene expression profile enabled our machine learning framework to simultaneously take all 53 tissues into account.

Applying the Bray-Curtis similarity measure to quantify the similarity in the expression profile between two genes is another advance of our machine learning framework. An intuitive comparison in the visualized overall expression pattern across the 53 tissues between NR3C1 and the two genes (i.e. SLC25A32 and TANK) with the highest similarity values computationally identified by the Bray-Curtis metric proved its efficacy. The Bray-Curtis similarity measure has three desired mathematical properties described in Chapter 1 that potently justify its application in this situation. The first is the strict bounds of 0 and 1 that it maintains, which form a natural similarity range that is easy to interpret. The second is the similarity between the expression profiles of the two genes can achieve the maximum value 1, if and only two genes exhibit the same expression value in each of the 53 tissues. This is attributable to the simultaneous consideration of both the directions and lengths of vectors by the Bray-Curtis function, compared to the Cosine similarity. In measurement of similarity between tissue-wide expression profiles, not only do the relative magnitudes of expression values of the gene across multiple tissues (i.e. the direction of the vector) matter, but the absolute expression value of the gene in every single tissue (i.e. the length of the vector) also matters. The third is the domination of the resulting similarity value by higher expression values. Intuitively, the proximity between larger expression values is more important in determining the similarity between the expression profiles of two genes than that between smaller expression values.

Another advance of the machine learning framework is its unifying capability to predict both gene expression profiles and TF target genes using the same set of features derived from heterotypic or homotypic TFBS clusters. These features captured the spatial distribution (e.g. the distances between clusters and the TSS) and information composition (e.g. the R_i values of binding sites) of clusters in the promoter. The rationale that the organization and composition of *cis*-regulatory modules in the promoter dictate gene expression patterns and identify correct TF targets confers this unifying capability.

The most direct generalization of the Bray-Curtis similarity measure is that it can be applied on any gene whose expression profile is available from GTEx to identify genes with overall similar expression patterns across 53 tissues to it. Furthermore, the definition of the gene expression profile is flexible; if another project other than GTEx also generated the expression data of genes across multiple tissues, the gene expression profile can be similarly redefined and the Bray-Curtis similarity measure is still usable. Another generalization of the Bray-Curtis similarity metric is to measure the similarity in the expression profile between different transcripts of the same gene.

In Chapter 6, the machine learning framework was applied to the TF knockdown data respectively in the forms of P-values and regulatory coefficients in the GM19238 and K562 cell lines generated by Cusanovich et al. (23) and Dixit et al (24). To date ENCODE has conducted 127 CRISPR- and 54 siRNA-based knockdown experiments for more TFs in the K562 cell line. Apart from raw RNA-seq reads data, ENCODE also directly provides the absolute gene expression values before and after TF knockdown, instead of indirect P-values and regulatory coefficients generated by the computational models. Thus for the machine learning framework to be applied to these newly generated ENCODE data, two preconditions need to be met. The first is that we need to derive iPWMs exhibiting primary binding motifs from ChIP-seq data for the sequence-specific TFs, enabling us to detect their binding sites. The second is that more flexibility needs to be added to the initial interface of the framework responsible for identifying DE TF targets from knockdown data, by adapting it to the different form of knockdown data (i.e. absolute gene expression values).

7.2 Implications of the Results Obtained in this Thesis

7.2.1 Transcription Factor Binding Site Motifs

The iPWMs exhibiting primary binding motifs of sequence-specific TFs can be used to detect binding sites and predict the effects of naturally occurring or artificially introduced variants on binding site strengths, as proved by the accurate detection of true binding sites and interpretation of effects of experimentally characterized SNPs in Chapter 3, prioritization of TFBS variants in HBOC genes in Chapters 4 and 5, and analysis on the

artificial mutations introduced into EGR1 binding sites in the MCM7 promoter and SNPs in the target genes of seven TFs in Chapter 6.

They can also be used to perform mutation analyses on all known SNPs present in the dbSNP database by first detecting whether they lie within TFBSs. In fact, these iPWMs have been integrated into MutationForecaster, a commercial web-based software suite that provides a single easy-to-navigate interface, interprets mutations that affect transcription, mRNA splicing and protein coding, and stores the results in a database that can be compared to other sources of genome variation. Another higher-level use of these iPWMs is to predict TF target genes, as demonstrated by the successful identification of DE direct targets of 7 and 11 TFs respectively in the GM19238 and K562 cell lines based on iPWM-detected binding sites and IDBC-detected clusters in Chapter 6.

Cofactor motifs are a systemic component of ChIP-seq datasets (25); for example, the CTCF motif frequently is significantly enriched in ChIP-seq peaks of other primary TFs, consistent with our finding that CTCF is revealed as a cofactor of SMC3, RAD21 and ZNF143. Peaks only containing CTCF motifs and lacking primary TFs' motifs compose up to 45% of a ChIP-seq dataset (25). These novel cofactor motifs enabled us to predict new TF-TF interactions and complexes. Future experimental studies can confirm the presence of these predicted complexes, ascertain the underlying physical mechanisms of these interactions, and investigate functional significance of these complexes (e.g. the biological pathways that they participate in). They can also confirm that the novel motifs are indeed functional, independent and can be recognized by TFs, rather than being general noise.

The list of experimentally confirmed TFBSs and the list of experimentally characterized SNPs that we compiled from the literature in Chapter 3 can serve as standard datasets to benchmark the accuracy of TFBS motifs and PWMs that future studies will derive from ChIP-seq datasets. As experimental studies identify more binding sites and the effects of more TFBS variants, the two lists will be maintained and expanded by incorporating these new findings, in terms of the numbers of binding sites, SNPs and different TFs.

7.2.2 Transcription Factor Binding Site and Other Variants in the Hereditary Breast and Ovarian Cancer Genes

Future experimental studies can confirm the predicted effects of the TFBS and other types of non-coding and coding variants that were prioritized in 20 complete genes of HBOC patients in Chapters 4 and 5 on binding site strengths, mRNA splicing, mRNA secondary structure and amino acid chains of the final proteins. For TFBS variants, they can also investigate whether and how the expression levels of these genes are altered, and whether the expression alterations of these genes are associated with HBOC, establishing the association of these variants with the increased risks of HBOC. On a higher level, they can further ascertain the specific functional pathways through which changes in the concentration and structure of the protein molecules encoded by these genes result in the onset of HBOC, after statistical studies prove that these variants are associated with the increased risks of HBOC.

7.2.3 Genes with Similar Tissue-wide Expression Profiles to, and Differentially Expressed, Direct Target Genes of the Transcription Factors

In Chapter 6, the genome-wide genes with overall similar expression patterns across 53 tissues to 17 TFs were identified by the Bray-Curtis similarity measure, and the general machine learning framework partly answered the question of why these genes exhibit similar tissue-wide expression profiles to the TFs by finding that the similarity in the distribution and composition of TFBS clusters in the proximal promoters partly explained the similarity in the expression profiles. Since the expression pattern of a gene is determined by both TF binding and epigenetic markers within all regulatory sequences including proximal promoters and distal enhancers/silencers, future studies can explore the similarity between these genes and the TFs in histone modification patterns within proximal promoters, and in other regulatory regions both in terms of TFBSs and histone modifications. In addition, the high similarity in the overall expression pattern across 53 tissues suggests the potential functional relatedness of the TFs to these genes, especially the subset of genes that are also DE direct targets of the TFs at the same time; for example, they may serve as key components within the same genomic pathway.

In Chapter 6, TF-cofactor co-regulation was found to be responsible for distinguishing functional binding sites in DE targets from non-functional binding sites in non-targets, which was proved by different flanking cofactor binding sites present in the promoters of targets and non-targets of YY1 and EGR1. Similar to the discovery of novel cofactor motifs in Chapter 3, future experimental studies can focus on the interactions between these cofactors (i.e. SP1, KLF1, CEBPB, NFY determining functional sites and CTCF, ETV6 determining non-functional sites) and the primary TFs (i.e. YY1 and EGR1), and how they affect the functional status of the binding sites.

The protection of multiple information-dense clusters in the promoter against TFBS mutations suggests that it is not sufficient for subsequent studies to predict only the effect of a mutation on the binding site strength; further, they also need to interpret the more downstream effects on the cluster and potentially gene expression. TFBS mutations that can be compensated for by other clusters may lower the promoter activity despite being still able to induce gene expression, leading to natural phenotypic diversity. The majority of previous studies focus on truly deleterious mutations leading to the onset of disease phenotypes; future studies can attempt to relate more mutations that can be compensated for to diverse non-disease phenotypes.

7.3 Potential Limitations and Future Studies

7.3.1 ChIP-seq datasets from which Maskminent Only Returned Noise Motifs

Among all ChIP-seq datasets analyzed in Chapter 3, there were ~20 datasets from which the Maskminent motif discovery pipeline was not able to discover the primary TF or any cofactor motif, and only returned noise motifs. There are two primary reasons explaining why these datasets are present, depending on whether the peaks contained in these datasets were pulled down due to binding by protein molecules (i.e. primary TFs, cofactors and histones) or by antibodies.

The majority of these datasets belong to non-sequence-specific TFs that primarily serve as catalyzing enzymes of histone modifications apart from acting as transcriptional cofactors interacting with DNA-binding TFs. For example, HDAC2 is responsible for the

deacetylation of lysine residues on the N-terminal region of the core histones (26) while being capable of forming transcription-repressing complexes with many DNA-binding TFs (e.g. YY1 (27) and SP1 (28)). And EZH2, as a subunit of PRC2 (Polycomb repressive complex 2), catalyzes trimethylation of H3K27 (H3K27me3) resulting in silent chromatin and eventually transcriptional repression (29), whereas it also can activate transcription via physical association with NFKB (30).

The functional versatility of these TFs explains why Maskminent did not detect any known TFBS motif from these datasets. Peaks contained in such a dataset can be classified into two categories, depending on the function that the primary TF was serving when they were pulled down. A larger number of peaks were bound by histone proteins that were bound by the primary TF that was catalyzing histone modifications, whereas histones bind to DNA sequences in a non-sequence-specific way; thus these peaks do not contain specific sequence motifs directly recognized by TFs. By contrast, the remaining smaller number of peaks were bound by sequence-specific cofactors with which the primary TF was associating, so that the binding sites recognized by the cofactor were enriched in these peaks. For example, in Chapter 3 Maskminent returned a low-complexity noise motif from the HepG2 dataset of HDAC2 consisting of 18,836 peaks, but detected FOXA motif from the top 5,000 peaks.

The other reason is that the peaks in such a dataset were isolated due to direct, non-specific binding by antibodies, rather than binding by protein molecules; thus these peaks are known as “noise” in the ChIP-seq technology, in contrast with the “signal” peaks that are pulled down in TF-DNA complexes and contain TFBSs. Sources that can introduce noise or bias into a ChIP-seq experiment include antibody quality, sequencing depth, library complexity, ChIP enrichment, differential protection against sonication across the genome, and differential mappability of short reads to repeat-rich genomic regions (31-33).

In fact, the peak calling step of a ChIP-seq experiment aims to achieve maximal signal-noise ratio when identifying peak intervals based on the enrichment level of isolated DNA fragments (or tags) (34). Prior studies developed multiple discriminative peak

Table 7.1: Peak calling algorithms

Algorithm	Background tag distribution	Operating process	Reference
SPP	Poisson	three binding detection algorithms to take advantage of the strand-specific tag pattern	(36)
CCAT	Poisson	distinguishing background from signal by maximizing signal-to-noise ratio between immunoprecipitation data and negative control	(35,37)
PeakSeq	Binomial	scoring signal sites relative to control under the null hypothesis of a binomial distribution of tags with a mean estimated from the number of tags in the negative control at the same site	(35,38)
BayesPeak	Binomial	using a negative binomial regression model, formulated as a Poisson-Gamma mixture, with parameters estimated from the negative control via Monte Carlo Markov chain methods	(35,39)
MACS	Poisson	using a variable rate Poisson model, where the model mean is determined from the negative control by taking the maximum of average read counts computed on 1kb, 5kb, 10kb, and genome-wide intervals	(35,40)

MOSAiCS	Binomial	using a negative binomial regression model computed from GC content, mappability, and a monomial in tag counts of the negative control, with a piecewise-defined mean	(35,41)
Coda	N/A	using convolutional neural networks to learn a generalizable mapping between ‘suboptimal’ and high-quality ChIP-seq data, while attenuating three primary sources of noise—due to low sequencing depth, low cell input and low ChIP enrichment	(42)

calling algorithms that distinguish signal from background noise in immunoprecipitation data based on a negative control into which the TF proteins of interest were not added (Table 7.1). Most of these algorithms assume that background noise tags conform to a null mathematical distribution (Poisson or Binomial), and the parameters of the distribution are estimated from the negative control and scaled to fit true immunoprecipitation data (35).

7.3.2 Predicted False Positive Transcription Factor Binding Sites

Traditional PWMs, which assume that individual positions within the binding site are independent of each other, tend to predict false positives (e.g. low-complexity sequences) when detecting TFBSs (43). However, position interdependencies within the binding site were confirmed to be present by the total mutual information of the iPWMs derived in Chapter 3 (Appendix B.2) and a variety of experimental techniques (44), including crystal structure analyses (45), quantitative multiple fluorescence relative affinity (QuMFRA) assays (46) and PBMs (47,48). PBM data have also demonstrated that position dependencies are stronger between neighboring positions than others (44,47,48). Thus one possible reason why traditional PWMs predict false positives is the underlying assumption of independence between individual positions within the binding site.

Note that these false positive binding sites predicted by PWMs differ from the non-functional binding sites distinguished by TF-cofactor coregulations from functional sites in Chapter 6. False positive sites are non-sites that cannot actually be recognized or bound by TFs (i.e. in the case of iPWMs, their ground-truth R_i values are < 0, but predicted R_i values are > 0); by contrast, both non-functional and functional sites are true binding sites that can be physically bound by TFs (i.e. their ground-truth R_i values are $>$ 0), with non-functional sites being unable to alter the gene expression level upon binding.

Apart from different types of interacting cofactors determining the functional state of a binding site, multiple other approaches that can enrich for functional sites and result in accurate feature sets were also simultaneously applied in Chapter 6; by contrast, they were individually used in prior studies, including focusing on proximal promoters (49), using DNase I hypersensitive data (50) and detecting information-dense clusters (51,52).

On the other hand, there are two additional approaches that were not incorporated into Chapter 6, using histone modification data (50) and focusing on phylogenetically conserved sequences across species (53).

7.3.2.1 Approaches to Reduce Predicted False Positive Binding Sites by Modelling Position Interdependencies

Prior studies applied three mainstream mathematical approaches to enable PWMs (e.g. frequency matrices) to capture position interdependencies within the TFBS and reduce the number of predicted false positive binding sites (Table 7.2), including Hidden Markov Models (HMMs) (44,47,54,55), straightforward oligonucleotide frequency matrices (56-58), and Bayesian networks (59-61). Other approaches include directly adding additional terms representing extra energy dissipated from the appearance of specific dinucleotides at pairwise positions into the equation to compute the energy dissipation of a TFBS from a binding energy-based PWM (62), and mixing frequency matrices computed from all binding sites and individual sites using a variable mixing parameter and pseudocount (63). These more complicated PWMs incorporating position interdependencies did achieve a higher specificity in detecting TFBSs; however, the improvement for most TFs is minor (44,62).

First-order HMMs can capture adjacent dinucleotide interdependencies: each position within the binding site corresponds to four internal states each of which emits one base with certainty, and a background state describing the nucleotide frequencies of flanking sequences is also present. The transition from the background state to an internal state represents the start of a binding site, and the transition probability between two internal states is the frequency of the dinucleotide appearing at the two positions (44,47). However, HMMs are less scalable; modelling interdependencies among multiple positions requires an increase in the order of HMMs, resulting in an exponential increase in the size and complexity of HMMs (55). Similarly, oligonucleotide frequency matrices also have a limited scalability; for example, modelling trinucleotide interdependencies implies that the frequencies of $4^3 = 256$ 3-mers need to be computed for every three positions. Thus only one prior study attempted to select a subset of k-mers ($k \geq 3$) based

Table 7.2: Approaches to model position interdependencies in PWMs to reduce predicted false positive binding sites

Approach	Mathematical model	Operating process	Reference
TFFMs (TF flexible models), Bulyk et al., PVLMM (permutated variable-length Markov model), FMM (feature motif model)	Hidden Markov Models (HMMs) and Markov networks	1st-order HMMs were used to model dinucleotide interdependencies and variable motif lengths. Each position corresponded to four states (i.e. four bases occurring at the position), and the transition probability between two states was derived from the frequency of the dinucleotide occurring at the two positions. To model interdependencies between non-adjacent positions, PVLMM searched for the best permutation of the motif positions. FMM used log-linear representations of Markov networks to model position interdependencies.	(44,47,54,55)
Ponomarenko et al., Stormo et al., Zhou et al.	Oligonucleotide alphabets/PWMs, dinucleotide PWMs	Based on thermodynamic, conformational and electrostatic properties of adjacent bases, a subset of oligonucleotides of each different length was contained in an alphabet. Then the frequency of each oligonucleotide was derived to form a frequency matrix in Ponomarenko et al. In Stormo et al. and Zhou et al. position frequency matrices contain the frequencies of dinucleotides occurring at adjacent pairwise positions. Zhou	(56-58)

		et al. restricts that any position can at most correlate with one other position.	
Barash et al., Ben-Gal et al., Pudimat et al.,	Bayesian networks	Applying the Bayesian Theorem to model position interdependencies (e.g. the probability of a dinucleotide occurring at two positions is the probability of the 1st base occurring at the 1st position, times the probability of the 2nd base occurring at the 2nd position under the condition that the 1st base has occurred at the 1st position).	(59-61)
BEM (binding energy model)	Binding energy-based PWMs	A binding energy-based PWM indicates the heat energy dissipated by each base at each position during the binding process. The additional energy dissipation caused by the occurrence of the specific dinucleotide at the two positions was explicitly added into the equation to compute the energy of a binding site.	(62)
King et al.	A parametric mixture of frequency matrices	Frequency matrices were derived from all binding sites and each individual site. One parameter was the pseudocount used to compute frequency matrices, and the other was the mixing weight between the two matrices from all sites and from one single site.	(63)

on thermodynamic, conformational and electrostatic properties of adjacent bases (56), whereas other studies adopted dinucleotide frequency matrices (57,58).

A Bayesian network can be represented by a directed acyclic graph in which a node denotes a position in the binding site and an edge denotes a dependency between two positions (59-61). This dependency is naturally modelled by the conditional probability in the Bayes' Theorem, which is used to compute the probability of a DNA sequence being a binding site (Equation 7.1).

$$P(N_1, N_2) = P(N_1) \cdot P(N_2|N_1) = P(N_2) \cdot P(N_1|N_2) \quad [7.1]$$

where N_1 and N_2 are respectively the nucleotides appearing at Positions 1 and 2. Equation 7.1 computes the probability of this dinucleotide appearing at the two positions in the presence of an interdependency between Position 1 and Position 2.

7.3.2.2 High-dimensional iPWMs to Model Position Interdependencies

Alternatively, we propose to use high-dimensional iPWMs, a natural generalization of the current two-dimensional (2D) iPWMs, to capture position interdependencies within the TFBS. For example, a 3D iPWM is capable of capturing dinucleotide interdependencies; the x and y dimensions represent individual positions, and the z dimension represents the sequence conservation of each dinucleotide appearing at the x and y positions, computed from its frequency and measured in bits of information.

From a multiple alignment of n binding sites, the frequencies of 16 dinucleotides at each pair of positions are determined. A contiguous 3D iPWM will be computed from

$$R_{iw}(dn, l_1, l_2) = 4 - \left(-\log_2 f(dn, l_1, l_2) + e\big(n(l_1, l_2)\big)\right) \ (bits\ per\ base) \quad [7.2]$$

where $f(dn, l_1, l_2)$ is the frequency of dinucleotide dn at positions l_1 and l_2 in the aligned binding site sequences and $e\big(n(l_1, l_2)\big)$ is a sampling error correction factor (64) at positions l_1 and l_2 for the n sequences used to create $f(dn, l_1, l_2)$.

Similarly, Equation 1.3 will also be generalized to compute the R_i value of a contiguous binding site j, which is the dot product between the sequence and the 3D iPWM.

$$R_i(j) = \sum_{l_1,l_2} \sum_{b=A}^{T} s(dn, l_1, l_2, j) R_{iw}(dn, l_1, l_2) \ (bits\ per\ site) \quad [7.3]$$

where $s(dn, l_1, l_2, j)$ is a 3D binary matrix for the sequence j, in which cells have a value of 1 for dinucleotide dn at position l_1 and l_2, and a value of 0 elsewhere. Assuming dinucleotide interdependencies only occur within the half site and do not occur across two half sites, the calculation of the R_i value of a bipartite binding site j will remain the same, except that the R_i value of either half site is computed from a 3D iPWM using Equation 7.2.

It can be anticipated that the unique binding site sequences predicted by the 3D iPWM will be a subset of those predicted by the 2D iPWM (Figure 7.1, Example 7.1), resulting in a smaller number of false positives and true positives (i.e. missing more true binding sites) at the same time.

Similarly, a 4D iPWM can model trinucleotide interdependencies, etc. The highest possible dimension is the binding site length, which can simultaneously capture interdependencies among all positions. The minimum set of unique binding site sequences detected by such an iPWM is identical to that contained in the multiple alignment used to compute the iPWM.

Compared to the aforementioned approaches modelling position interdependencies, high-dimensional iPWMs can simultaneously capture the interdependencies among all possible combinations of positions. For example, the widely used 1st-order HMM only naturally captures adjacent dinucleotide dependencies, since taking into account non-adjacent positions will significantly increase its complexity. By contrast, a 3D iPWM automatically incorporates the interactions between all pairwise positions. However, further studies on experimentally confirmed non-sites are needed to prove that high-dimensional iPWMs are indeed capable of predicting less false positives, which is beyond the scope of this thesis.

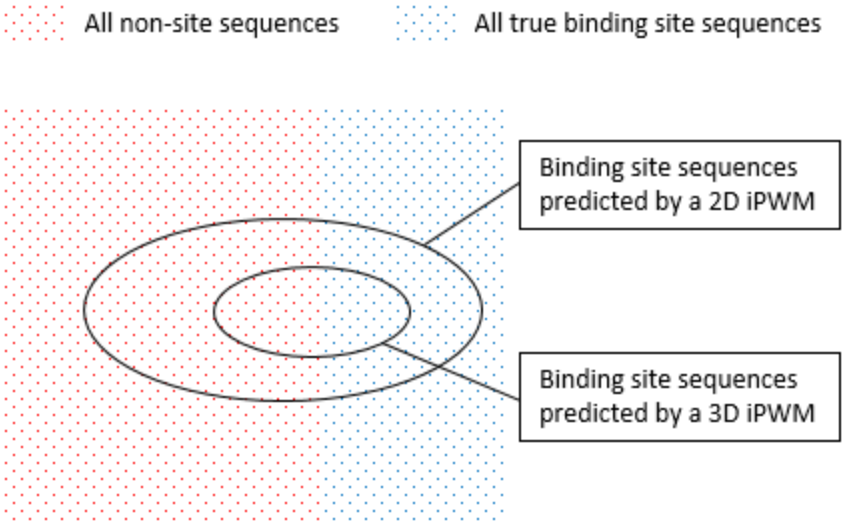

Figure 7.1: Unique binding site sequences predicted by a 2D iPWM and a 3D iPWM. The unique binding site sequences predicted by an iPWM include true sites, as well as non-sites. Using a 3D iPWM will simultaneously increase specificity and decrease sensitivity compared to a 2D one.

Example 7.1: Suppose that TATA, TAAT, TAGC, TACG are four true binding sites of some TF. According to Equations 1.1 and 1.3, the information content of the DNA sequence TAAA computed from the 2D iPWM is $R_i(TAAA) = 4\ bits$ (for simplicity, the pseudocount and sampling error correction factor are not taken into account). Thus TAAA will be predicted to be a binding site. On the other hand, according to Equation 7.2 the 3D iPWM derived from the alignment of the four binding sites is shown in Table 7.3.

According to Equation 7.3, the information content of the DNA sequence TAAA computed from the 3D iPWM is $R_i(TAAA) = -\infty$ bits. Thus it will not be predicted to be a binding site. Since TAAA is a lower-complexity sequence compared to the four true sites and the dinucleotide AA never appears at Positions 3 and 4, it is more likely to be a non-site, which justifies the prediction made by the 3D iPWM.

Table 7.3: The 3D iPWM in Example 7.1

Dinucleotide	Pairwise positions					
	1,2	**1,3**	**1,4**	**2,3**	**2,4**	**3,4**
AA	−∞	−∞	−∞	2	2	−∞
AT	−∞	−∞	−∞	2	2	2
AC	−∞	−∞	−∞	2	2	−∞
AG	−∞	−∞	−∞	2	2	−∞
TA	4	2	2	−∞	−∞	2
TT	4	2	2	−∞	−∞	−∞
TC	4	2	2	−∞	−∞	−∞
TG	4	2	2	−∞	−∞	−∞
CA	−∞	−∞	−∞	−∞	−∞	−∞
CT	−∞	−∞	−∞	−∞	−∞	−∞
CC	−∞	−∞	−∞	−∞	−∞	−∞
CG	−∞	−∞	−∞	−∞	−∞	2
GA	−∞	−∞	−∞	−∞	−∞	−∞
GT	−∞	−∞	−∞	−∞	−∞	−∞
GC	−∞	−∞	−∞	−∞	−∞	2
GG	−∞	−∞	−∞	−∞	−∞	−∞

7.3.3 Downstream Effects of Transcription Factor Binding Site Variants and Long-range DNA Interactions

As described in Chapter 6, the ideal prediction of effects of a TFBS variant of uncertain significance ought to include three levels, the binding site strength, binding site cluster, and gene expression. In Chapters 4 and 5, the unified framework prioritized a number of variants within binding sites of multiple TFs (e.g. CEBPB, HSF1) known to play a role in

breast cancer development in the HBOC genes. However, this framework only predicted the effects of these variants on binding site strengths, since this study had been conducted earlier.

Thus a future study will extend the framework by integrating predictions of effects of TFBS variants on information-dense clusters and gene expression, and apply this extended framework to reanalyze the prioritized variants in Chapters 4 and 5. This will further prioritize these variants by classifying them into two categories, truly deleterious mutations that can actually significantly alter gene expression and result in disease phenotypes, and variants that can be compensated for by other clusters and result in diverse natural phenotypes.

In fact, deleterious variants in the binding sites of many TFs have been known to be associated with a variety of diseases (Table 7.4), which in turn reflect the tissue-specific functions of the TFs. For example, 797 established SNPs associating with 144 diseases were found to lie within binding sites of NFKB, a significant overrepresentation (2.25-fold) compared with all common variants (P-value = 4.2×10^{-90}) (65). SNPs associated with primarily inflammatory and autoimmune diseases and cancers, including rheumatoid arthritis, systemic lupus erythematosus, primary biliary cirrhosis, asthma, and lymphoma, were highly enriched in NFKB binding sites (65). This is consistent with the known NFKB-mediated regulation of various aspects of innate and adaptive immune responses, including the transcription of cytokines and antimicrobial effectors, and the development and survival of the cells and tissues that carry out immune responses (66). Similarly, the B cell-specific EBF1 is essential for maintaining B cell identity and preventing alternative fates in committed cells (76), which accounts for the correlation between SNPs resulting in allele-specific binding of EBF1 and autoimmune diseases (77-80).

A recent study used the Hi-C technology to identify TF-mediated long-range interactions between 31,253 promoters and distant regulatory elements (e.g. distal enhancers and promoters of other genes) in 17 human primary hematopoietic cell types (81). 698,187 high-confidence unique promoter interactions were detected across all cell types, of which 9.6% were promoter-to-promoter interactions and 90.4% promoter-to-enhancers,

Table 7.4: Association of TFBS variants with diseases

TF	Diseases with which SNPs in the binding sites of the TF are associated	Functions of the TF in the specific tissue	Reference
NFKB	Inflammatory and autoimmune diseases and cancers (e.g. asthma)	Regulates various aspects of innate and adaptive immune responses	(65,66)
CTCF	Breast cancer (BC) (e.g. rs11540855)	Affects BC development by regulating target genes; the hypopoly(ADPribosyl)ated isoform is characteristic for BC	(67-69)
	Skin cancer	CTCF knockdown promoted invasion, metastasis and epithelial-mesenchymal transition in liver and skin cancer.	(70,71)
	Liver cancer		(71,72)
	Gastrointestinal Cancers	Associations were established between motif mutations and late replication.	(70,71)
GABPA	BC (e.g. rs2853669)	Controls cell migration in breast epithelial cells	(67,73,74)
USF1/2	BC (e.g. rs3760982)	Have a potent growth-inhibitory effect and loss of USF function favors cell proliferation.	(67,75)
EBF1	Autoimmune diseases (psoriasis, primary biliary cirrhosis and rheumatoid arthritis) (e.g. rs909685, rs9603612)	Essential for the maintenance of B cell identity and prevention of alternative fates in committed cells.	(76-80)

with a median linear distance between promoters and their interacting regions of 331 kb (81). These promoter interactions were found to be highly cell type specific, and interacting regions are enriched in genetic variants linked with altered expression of genes they contact (81). However, this study did not identify which specific TFBSs present in the promoters and distal interacting regions are responsible for the formation of long-range loops. As described above, this unified framework can be directly generalized to analyze all variants in the whole genome instead of specific genes. Therefore, this unified framework can be further extended by first ascertaining whether the prioritized TFBS variants lie in distal interacting regions using the Hi-C data, then predicting how these variants affect the expression levels of the distant genes whose promoters form long-range loops with the interacting regions. For example, the intron 19 of the CLEC16A gene serves as a distal enhancer (~160 kb away) interacting with the promoter of DEXI gene (82), and the SNP (rs12708716) within this intron associated with the type 1 diabetes significantly altered the expression level of DEXI (82).

The Hi-C long-range interacting data can also be used to improve the general machine learning framework for prediction of genes with similar tissue-wide expression profiles in Chapter 6. After the Bray-Curtis similarity measure identify the genes with similar expression profiles to the TFs, distal interacting regions that can perform long-range interactions with the promoters of these genes can be further obtained using the Hi-C data. The same feature sets will be derived from the information-dense clusters detected from iPWM-detected TFBSs within these interacting regions by the IDBC algorithm. Since the expression pattern of a gene is determined by all regulatory regions, the high similarity between these genes and the TFs in the tissue-wide expression profile is attributable to the high similarity in all regulatory regions including long-range interactions. Therefore, the incorporation of the spatial organization and information composition of transcriptional regulatory modules in the distal interacting regions into the machine learning framework, in addition to those in the proximal promoters, will result in an improvement in the classifier performance.

7.4 Conclusions

Compared to prior studies, this thesis presents an improved three-level computational modelling of the transcriptional regulation of human genes, involving TFBSs, information-dense clusters, and gene expression. The ultimate goal of transcriptional regulation is to accurately regulate expression levels of TF target genes via the underlying physical interactions between TFs and individual binding sites.

The lowest level is the derivation of contiguous and bipartite iPWMs from ENCODE ChIP-seq datasets modelling TF binding specificities in Chapter 3. The information content of a binding site computed from an iPWM is quantitatively related to the amount of heat energy dissipated by the TF-binding site physical association, enabling iPWMs to more accurately quantify binding site strengths than log likelihood ratio-based PWMs derived by prior studies. The bipartite iPWMs more precisely model the binding behavior of dimeric TFs by taking into account the variable-length spacers within bipartite binding sites. Compared to prior studies only analyzing a small number of top peaks, the derived iPWMs incorporated the maximum number of intermediate and weak binding sites via the recursive thresholding functionality. This enabled the accurate quantification of binding site strengths across a broad range of affinities, which was proven by the successful detection of true binding sites and interpretation of experimentally measured effects of SNPs in Chapter 3, and prioritization of TFBS variants in HBOC genes in Chapters 4 and 5.

The intermediate level is the relationship between individual binding sites and clusters in terms of composition and variation in Chapter 6. Compared to prior clustering algorithms, the detection of the information-dense clusters by the IDBC algorithm simultaneously rely on both the spatial distribution and information contents of binding sites, enabling the more accurate modelling of the clustering composition of binding sites. Apart from the additive, complementary cooperation between individual sites within a cluster on inducing gene expression found by prior studies, mutation analyses on artificially introduced mutations and naturally occurring SNPs also revealed another compensatory cooperation; that is, the presence of multiple binding sites in a cluster

enable the cluster to be robust against mutations by compensating for each other's destruction, with the central strong site playing a more significant role.

The highest level is the relationship between individual clusters and gene expression in terms of composition and variation in Chapter 6. The Bray-Curtis similarity measure and gene knockdown data respectively enabled the accurate identification of similar tissue-wide gene expression profiles and differentially expressed TF target genes. Machine learning features accurately modeled the spatial organization and information composition of TFBS clusters in proximal promoters, which substantially dictate the expression profiles of TF target genes. Mutation analyses on TF targets revealed that the presence of multiple information-dense clusters in a promoter enable gene expression to be robust against TFBS mutations by compensating for each other's destruction, relating deleterious and protective variants respectively to disease and diverse natural phenotypes.

Therefore, by comprehensively delineating physical TF-binding site interactions, functional binding site-binding site interactions within the information-dense cluster and cluster-cluster interactions within the promoter, this thesis aims to improve the current computational modelling of human transcriptional regulation.

7.5 References

1. Bailey,T.L., Williams,N., Misleh,C. and Li,W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res., 34, W369–W373.

2. Wang,J., Zhuang,J., Iyer,S., Lin,X., Whitfield,T.W., Greven,M.C., Pierce,B.G., Dong,X., Kundaje,A., Cheng,Y., et al. (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res., 22, 1798–1812.

3. Kheradpour,P. and Kellis,M. (2014) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res., 42, 2976–2987.

4. Setty,M. and Leslie,C.S. (2015) SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps. PLoS Comput. Biol., 11, e1004271.

5. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K., et al. (2014) Determination and inference of eukaryotic transcription factor sequence specificity. Cell, 158, 1431–1443.

6. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G., et al. (2013) DNA-binding specificities of human transcription factors. Cell, 152, 327–339.

7. Toivonen,J., Kivioja,T., Jolma,A., Yin,Y., Taipale,J. and Ukkonen,E. Modular discovery of monomeric and dimeric transcription factor binding motifs for large data sets. Nucleic Acids Res., 10.1093/nar/gky027.

8. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. Nat. Rev. Genet., 10, 252–263.

9. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The Human Transcription Factors. Cell, 172, 650–665.

10. Caminsky,N., Mucaki,E.J. and Rogan,P.K. (2014) Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis. F1000Research, 3, 282.

11. Peterlongo,P., Catucci,I., Colombo,M., Caleca,L., Mucaki,E., Bogliolo,M., Marin,M., Damiola,F., Bernard,L., Pensotti,V., et al. (2015) FANCM c.5791C>T nonsense mutation (rs144567652) induces exon skipping, affects DNA repair activity and is a familial breast cancer risk factor. Hum. Mol. Genet., 24, 5345–5355.

12. Mucaki,E.J., Shirley,B.C. and Rogan,P.K. (2013) Prediction of mutant mRNA splice isoforms by information theory-based exon definition. Hum. Mutat., 34, 557–565.

13. Olsen,R.K.J., Brøner,S., Sabaratnam,R., Doktor,T.K., Andersen,H.S., Bruun,G.H., Gahrn,B., Stenbroen,V., Olpin,S.E., Dobbie,A., et al. (2014) The ETFDH c.158A>G variation disrupts the balanced interplay of ESE- and ESS-binding proteins thereby causing missplicing and multiple Acyl-CoA dehydrogenation deficiency. Hum. Mutat., 35, 86–95.

14. Ye,Z., Song,H., Higgins,J.P.T., Pharoah,P. and Danesh,J. (2006) Five glutathione s-transferase gene variants in 23,452 cases of lung cancer and 30,397 controls: meta-analysis of 130 studies. PLoS Med., 3, e91.

15. Song,N., Shin,A., Park,J.W., Kim,J. and Oh,J.H. (2017) Common risk variants for colorectal cancer: an evaluation of associations with age at cancer onset. Sci. Rep., 7, 40644.

16. Hart,T., Chandrashekhar,M., Aregger,M., Steinhart,Z., Brown,K.R., MacLeod,G., Mis,M., Zimmermann,M., Fradet-Turcotte,A., Sun,S., et al. (2015) High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell, 163, 1515–1526.

17. Hellman,M. (1980) A cryptanalytic time-memory trade-off. IEEE Trans. Inf. Theory, 26, 401–406.

18. Delisi,C. (1974) Cooperative phenomena in homopolymers. An alternative formulation of the partition function. Biopolymers, 13, 1511–1512.

19. Ouyang,Z., Zhou,Q. and Wong,W.H. (2009) ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proc. Natl. Acad. Sci. U. S. A., 106, 21521–21526.

20. Cheng,C., Alexander,R., Min,R., Leng,J., Yip,K.Y., Rozowsky,J., Yan,K.-K., Dong,X., Djebali,S., Ruan,Y., et al. (2012) Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome Res., 22, 1658–1667.

21. Budden,D.M., Hurley,D.G., Cursons,J., Markham,J.F., Davis,M.J. and Crampin,E.J. (2014) Predicting expression: the complementary power of histone modification and transcription factor binding data. Epigenetics Chromatin, 7, 36.

22. Smith,A.D., Sumazin,P., Xuan,Z. and Zhang,M.Q. (2006) DNA motifs in human and mouse proximal promoters predict tissue-specific expression. Proc. Natl. Acad. Sci. U. S. A., 103, 6275–6280.

23. Cusanovich,D.A., Pavlovic,B., Pritchard,J.K. and Gilad,Y. (2014) The functional consequences of variation in transcription factor binding. PLoS Genet., 10, e1004226.

24. Dixit,A., Parnas,O., Li,B., Chen,J., Fulco,C.P., Jerby-Arnon,L., Marjanovic,N.D., Dionne,D., Burks,T., Raychowdhury,R., et al. (2016) Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell, 167, 1853–1866.e17.

25. Worsley Hunt,R. and Wasserman,W.W. (2014) Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. Genome Biol., 15, 412.

26. Seto,E. and Yoshida,M. (2014) Erasers of Histone Acetylation: The Histone Deacetylase Enzymes. Cold Spring Harb. Perspect. Biol., 6.

27. Yao,Y.L., Yang,W.M. and Seto,E. (2001) Regulation of transcription factor YY1 by acetylation and deacetylation. Mol. Cell. Biol., 21, 5979–5991.

28. Won,J., Yim,J. and Kim,T.K. (2002) Sp1 and Sp3 recruit histone deacetylase to repress transcription of human telomerase reverse transcriptase (hTERT) promoter in normal human somatic cells. J. Biol. Chem., 277, 38230–38238.

29. Viré,E., Brenner,C., Deplus,R., Blanchon,L., Fraga,M., Didelot,C., Morey,L., Van Eynde,A., Bernard,D., Vanderwinden,J.-M., et al. (2006) The Polycomb group protein EZH2 directly controls DNA methylation. Nature, 439, 871–874.

30. Lee,S.T., Li,Z., Wu,Z., Aau,M., Guan,P., Karuturi,R.K.M., Liou,Y.C. and Yu,Q. (2011) Context-specific regulation of NF-κB target gene expression by EZH2 in breast cancers. Mol. Cell, 43, 798–810.

31. Jung,Y.L., Luquette L.J., Ho J.W.K., Ferrari F., Tolstorukov M., Minoda A., Issner R., Epstein C.B., Karpen G.H., Kuroda M.I., Park P.J. (2014) Impact of sequencing depth in ChIP-seq experiments. Nucleic Acids Res., 42, e74.

32. Teytelman L., Ozaydin B., Zill O., Lefrançois P., Snyder M., Rine J., Eisen M.B. (2009) Impact of chromatin structures on DNA processing for genomic analyses. PLoS One. 4, e6700.

33. Aird D., Ross M.G., Chen W.S., Danielsson M., Fennell T., Russ C., Jaffe D.B., Nusbaum C., Gnirke A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biology. 12, R18.

34. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P., et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res., 22, 1813–1831.

35. Diaz A., Park K., Lim D.A., Song J.S. (2012) Normalization, bias correction, and peak calling for ChIP-seq. Stat. Appl. Genet. Mol. Biol. 11(3).

36. Kharchenko P.V., Tolstorukov M.Y., Park P.J. (2008) Design and analysis of ChIP-seq experiments for DNAbinding proteins. Nat. Biotechnol. 26, 1351–9.

37. Xu H., Handoko L., Wei X., Ye C., Sheng J., Wei C.L., Lin F., Sung W.K. (2010) A signal-noise model for significance analysis of ChIP-seq with negative control. Bioinformatics. 26(9), 1199-204.

38. Rozowsky J., Euskirchen G., Auerbach R.K., Zhang Z.D., Gibson T., Bjornson R., Carriero N., Snyder M., Gerstein M.B. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat. Biotechnol. 27, 66–75.

39. Spyrou C., Stark R., Lynch A.G., Tavaré S. (2009) BayesPeak: Bayesian analysis of ChIP-seq data. BMC Bioinformatics. 10, 299.

40. Zhang Y., Liu T., Meyer C.A., Eeckhoute J., Johnson D.S., Bernstein B.E., Nussbaum C., Myers R.M., Brown M., Li W., Liu X.S. (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biology. 9, R137.

41. Kuan P.F., Chung D., Pan G., Thomson J.A., Stewart R., Keles S. (2011) A Statistical Framework for the Analysis of ChIP-Seq Data. J. Am. Sta.t Assoc. 106(495), 891-903.

42. Koh P.W., Pierson E., Kundaje A. (2017) Denoising genome-wide histone ChIP-seq with convolutional neural networks. Bioinformatics. 33, i225–i233.

43. Morozov,V.Y. and Ioshikhes,I.P. (2013) Optimized position weight matrices in prediction of novel putative binding sites for transcription factors in the Drosophila melanogaster genome. PloS One, 8, e68712.

44. Mathelier,A. and Wasserman,W.W. (2013) The Next Generation of Transcription Factor Binding Site Prediction. PLOS Comput. Biol., 9, e1003214.

45. Luscombe N.M., Laskowski R.A., Thornton J.M. (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res. 29, 2860–74.

46. Man T.K., Stormo G.D. (2001) Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. Nucleic acids research. 29, 2471–8.

47. Bulyk,M.L., Johnson,P.L.F. and Church,G.M. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res., 30, 1255–1261.

48. Berger M.F., Philippakis A.A., Qureshi A.M., He F.S., Estep P.W., et al. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat. Biotechnol. 24, 1429–35.

49. Bernard V., Lecharny A., Brunaud V. (2010) Improved detection of motifs with preferential location in promoters. Genome. 53, 739–52.

50. Arvey A., Agius P., Noble W.S., Leslie C. (2012) Sequence and chromatin determinants of cell-type-specific transcription factor binding. Genome Research. 22, 1723–34.

51. Ho Sui S.J., Mortimer J.R., Arenillas D.J., Brumm J., Walsh C.J., et al. (2005) oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. Nucleic Acids Res., 33, 3154–64.

52. Ho Sui S.J., Fulton D.L., Arenillas D.J., Kwon A.T., Wasserman W.W. (2007) oPOSSUM: integrated tools for analysis of regulatory motif over-representation. Nucleic Acids Res., 35, W245–52.

53. Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, et al. (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature., 450, 219–32.

54. Zhao X., Huang H., Speed T.P. (2005) Finding short DNA motifs using permuted Markov models. J Comput Biol., 12, 894–906.

55. Sharon E., Lubliner S., Segal E. (2008) A Feature-Based Approach to Modeling Protein–DNA Interactions. PLoS Comput Biol, 4(8): e1000154.

56. Ponomarenko,M.P., Ponomarenko,J.V., Frolov,A.S., Podkolodnaya,O.A., Vorobyev,D.G., Kolchanov,N.A. and Overton,G.C. (1999) Oligonucleotide frequency matrices addressed to recognizing functional DNA sites. Bioinformatics., 15, 631–643.

57. Stormo,G.D., Schneider,T.D. and Gold,L. (1986) Quantitative analysis of the relationship between nucleotide sequence and functional activity. Nucleic Acids Res., 14, 6661–6679.

58. Zhou Q., Liu J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions. Bioinformatics., 20(6), 909–916.

59. Barash,Y., Elidan,G., Friedman,N. and Kaplan,T. (2003) Modeling Dependencies in Protein-DNA Binding Sites. ACM, Berlin, pp 28–37

60. Ben-Gal I., Shani A., Gohr A., Grau J., Arviv S., et al. (2005) Identification of transcription factor binding sites with variable-order Bayesian networks. Bioinformatics 21, 2657–2666.

61. Pudimat R., Schukat-Talamazzini E.G., Backofen R. (2005) A multiple-feature framework for modelling and predicting transcription factor binding sites. Bioinformatics, 21, 3082–3088.

62. Zhao Y., Ruan S., Pandey M., Stormo G.D. (2012) Improved Models for Transcription Factor Binding Site Identifi cation Using Nonindependent Interactions. Genetics, 191, 781 –790.

63. King, O.D., Roth F.P. (2003) A non-parametric model for transcription factor binding sites. Nucleic Acids Res., 31, e116.

64. Schneider,T.D., Stormo,G.D., Gold,L., Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. J. Mol. Biol., 188, 415–431.

65. Karczewski K.J., Dudley J.T., Kukurba K.R., Chen R., Butte A.J., Montgomery S.B., Snyder M. Systematic functional regulatory assessment of disease-associated variants. Proc. Natl. Acad. Sci. 110 (23), 9607-9612.

66. Hayden M.S., West A.P., Ghosh S. (2006) NF-κB and the immune response. Oncogene. 25, 6758–6780.

67. Liu Y., Walavalkar N.M., Dozmorov M.G., Rich S.S., Civelek M., Guertin M.J. (2017) Identification of breast cancer associated variants that modulate transcription factor binding. PLoS Genet. 13(9):e1006761.

68. Oh S., Oh C., Yoo K.H. (2017) Functional roles of CTCF in breast cancer. BMB Rep. 50(9), 445-453

69. Docquier F., Kita G., Farrar D., Chernukhin I., Klenova E. (2008) Role of poly(ADPribosyl)ation of CTCF in cancer and normal breast cells. Breast Cancer Res. 10:P10.

70. Poulos R.C., Thoms J.A.I., Guan Y.F., Unnikrishnan A., Pimanda J.E., Wong J.W.H. (2016) Functional Mutations Form at CTCF-Cohesin Binding Sites in Melanoma Due to Uneven Nucleotide Excision Repair across the Motif. Cell Rep. 17(11), 2865-2872.

71. Wozniak R.J., Klimecki W.T., Lau S.S., Feinstein Y., Futscher B.W. 5-Aza-2′-deoxycytidine-mediated reductions in G9A histone methyltransferase and histone H3 K9 di-methylation levels are linked to tumor suppressor gene reactivation. Oncogene. 26:77–90.

72. Umer H.M., Cavalli M., Dabrowski M.J., Diamanti K., Kruczyk M., Pan G., Komorowski J., Wadelius C. (2016) A Significant Regulatory Mutation Burden at a High-Affinity Position of the CTCF Motif in Gastrointestinal Cancers. Hum Mutat. 37(9), 904-13.

73. Helbig S., Wockner L., Bouendeu A., Hille-Betz U., McCue K., French J.D., Edwards S.L., Pickett H.A., Reddel R.R., Chenevix-Trench G., et al. (2017) Functional dissection of breast cancer risk-associated TERT promoter variants. Oncotarget. 8, 67203-67217.

74. Odrowaz Z., Sharrocks A.D. (2012) The ETS Transcription Factors ELK1 and GABPA Regulate Different Gene Networks to Control MCF10A Breast Epithelial Cell Migration. PLoS One. 7(12):e49892.

75. Ismail P.M., Lu T., Sawadogo M. (1999) Loss of USF transcriptional activity in breast cancer cell lines. Oncogene. 18(40), 5582-91.

76. Nechanitzky R., Akbas D., Scherer S., Györy I., Hoyler T., Ramamoorthy S., Diefenbach A., Grosschedl R. (2013) Transcription factor EBF1 is essential for the maintenance of B cell identity and prevention of alternative fates in committed cells. Nat Immunol. 14(8), 867-75.

77. Cavalli M., Pan G., Nord H., Wallerman O., Wallén Arzt E., Berggren O., Elvers I., Eloranta M.L., Rönnblom L., Lindblad Toh K., Wadelius C. (2016) Allele-specific transcription factor binding to common and rare variants associated with disease and gene expression. Hum Genet. 135(5), 485-97.

78. Okada Y., Wu D., Trynka G., Raj T., Terao C., Ikari K., Kochi Y., Ohmura K., Suzuki A., Yoshida S., et al (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 506, 376–381.

79. Liu J.Z., Almarri M.A., Gaffney D.J., Mells G.F., Jostins L., Cordell H.J., Ducker S.J., Day D.B., Heneghan M.A., Neuberger J.M., et al. (2012) Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. Nat Genet. 44, 1137–1141.

80. Liu Y., Helms C., Liao W., Zaba L.C., Duan S., Gardner J., Wise C., Miner A., Malloy M.J., Pullinger C.R., et al. (2008) A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. PLoS Genet. 4:e1000041.

81. Javierre B.M., Burren O.S., Wilder S.P., Kreuzhuber R., Hill S.M., Sewitz S., Cairns J., Wingett S.W., Várnai C., Thiecke M.J. (2016) Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell. 167, 1369–1384.

82. Davison L.J., Wallace C., Cooper J.D., Cope N.F., Wilson N.K., Smyth D.J., Howson J.M., Saleh N., Al-Jeffery A., Angus K.L. (2012) Long-range DNA looping and gene expression analyses identify DEXI as an autoimmune disease candidate gene. Hum Mol Genet. 21(2), 322-33.

Curriculum Vitae

Name: Ruipeng Lu

Post-secondary Education and Degrees:

Shandong University
Jinan, Shandong, China
2007-2011 B.E.

Shandong University
Jinan, Shandong, China
2011-2014 M.E.

University of Luxembourg
Luxembourg, Luxembourg
2012-2014 M.Sc.

The University of Western Ontario
London, Ontario, Canada
2014-2018 Ph.D.

Honours and Awards: Western Graduate Research Scholarship
2014-2018

Related Work Experience Teaching Assistant
The University of Western Ontario
2014-2018

Publications:

Ruipeng Lu, Peter K. Rogan. (2018) Clustered, information-dense transcription factor binding sites identify genes with similar tissue-wide expression profiles. *bioRxiv*. doi: https://doi.org/10.1101/283267.

Ruipeng Lu, Eliseos Mucaki, Peter K. Rogan. (2017) Discovery and validation of information theory-based transcription factor and cofactor binding site motifs. *Nucleic Acids Research*, 45:e27

Eliseos J. Mucaki, Natasha G. Caminsky, Amy M. Perri, Ruipeng Lu, Alain Laederach, Matthew Halvorsen, Joan H. M. Knoll, Peter K. Rogan (2016) A unified analytic framework for prioritization of non-coding variants of uncertain significance in heritable breast and ovarian cancer. *BMC Med. Genomics*, 9, 19.

Natasha G. Caminsky, Eliseos J. Mucaki, Amy M. Perri, Ruipeng Lu, Joan H. M. Knoll, Peter K. Rogan (2016) Prioritizing Variants in Complete Hereditary Breast and Ovarian

Cancer (HBOC) Genes in Patients Lacking known BRCA Mutations. *Hum. Mutat.*, 10.1002/humu.22972

Xihui Chen, Piotr Kordy, Ruipeng Lu, Jun Pang. (2014) MinUS: Mining User Similarity with Trajectory Patterns. *ECML/PKDD* (3): 436-439.

Xihui Chen, Ruipeng Lu, Xiaoxing Ma, Jun Pang. (2014) Measuring User Similarity with Trajectory Patterns: Principles and New Metrics. *APWeb* 2014: 437-448