Improving the prediction of transcription factor binding sites to aid the interpretation of non-coding single nucleotide variants.

Narayan Jayaram

Research Department of Structural and Molecular Biology

University College London

A thesis submitted to University College London for the degree of Doctor of Philosophy

Declaration

I, Narayan Jayaram confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Narayan Jayaram

Abstract

Single nucleotide variants (SNVs) that occur in transcription factor binding sites (TFBSs) can disrupt the binding of transcription factors and alter gene expression which can cause inherited diseases and act as driver SNVs in cancer. The identification of SNVs in TFBSs has historically been challenging given the limited number of experimentally characterised TFBSs. The recent ENCODE project has resulted in the availability of ChIP-Seq data that provides genome wide sets of regions bound by transcription factors. These data have the potential to improve the identification of SNVs in TFBSs. However, as the ChIP-Seq data identify a broader range of DNA in which a transcription factor binds, computational prediction is required to identify the precise TFBS. Prediction of TFBSs involves scanning a DNA sequence with a Position Weight Matrix (PWM) using a pattern matching tool.

In this thesis, the prediction of TFBSs has been improved by: (a) evaluating a set of locally-installable pattern-matching tools and identifying the best performing tool (FIMO), (b) using the ENCODE ChIP-Seq data to evaluate a set of *de novo* motif discovery tools that are used to derive PWMs which can handle large volumes of data, (c) identifying the best performing tool (rGADEM), (d) using rGADEM to generate a set of PWMs from the ENCODE ChIP-Seq data and (e) by finally checking that the selection of the best pattern matching tool is not unduly influenced by the choice of PWMs.

These analyses were exploited to obtain a set of predicted TFBSs from the ENCODE ChIP-Seq data.

The predicted TFBSs were utilised for a comprehensive analysis of the Shannon entropy values of somatic cancer driver, and passenger SNVs that occur in TFBSs. Clear signals in Shannon entropy values were identified, and subsequently exploited to identify a threshold that can be used to prioritize driver SNVs for experimental validation.

Acknowledgements

First and foremost, I would like to thank my supervisor Dr Andrew Martin for his support and guidance throughout this PhD and for giving me this opportunity. I would also like to thank Dr Chris Taylorson, chair of my thesis committee and Prof Christine Orengo, member of my thesis committee for their feedback on my work. Thanks are also due to the UCL IMPACT Studentship scheme for funding.

I would also like to thank my consultant Dr Robin Lachmann, for his ongoing expert medical care, and encouragement at The National Hospital for Neurology and Neurosurgery, Prof James Leonard, my consultant during my childhood and early teenage years in London for his expert medical care at Great Ormond Street Hospital, and Prof Mark Batshaw who is now at the Childrens National Health System in Washington DC, for his immense help when I was a newborn in Scotland. Without them I would not be where I am today.

Contents

List of Figures

List of Tables

List of Abbreviations

ACCg	Geometric Accuracy
AN	Actual Negatives
bp	base pairs
BWT	Burrows-Wheeler Transform
ChIP	Chromatin ImmunoPrecipitation
DNA	DeoxyRiboNucleic Acid
EMSA	Electro-Mobility Shift Assay
FN	False Negatives
FP	False Positives
FPRs	False Positive Rate on Scrambled Sequences
GRC	Genome Reference Consortium
hCRM	Human ChIP-Seq rGADEM matrices
ICGC	International Cancer Genome Consortium
IUPAC	International Union of Pure and Applied Chemistry
MCC	Matthews Correlation Coefficient
PCR	Polymerase Chain Reaction
PWM	Position Weight Matrix
SAM	Sequence Alignment/Map format
SELEX	Systematic Evolution of Ligands by Exponential enrichment
Sn	Sensitivity
SNV	Single Nucleotide Variant
TCGA	The Cancer Genome Atlas
TFBS	Transcription Factor Binding Site
TFFM	Transcription Factor Flexible Models
TN	True Negatives
TP	True Positives
UCSC	University Of Santa Cruz

1 Introduction

This thesis is concerned with identifying and understanding the effects of Single Nucleotide Variants (SNVs) that occur in transcription factor binding sites (TFBSs). This is an area where little research has been carried out. This is in contrast to the large volume of research that has been carried out on investigating and predicting the effects of non-synonymous SNVs on protein structure and function.

1.1 Single Nucleotide Variants

SNVs are the most common form of genetic variation. SNVs are point mutations that result in base substitutions (Altshuler *et al.*, 2010; Cline and Karchin, 2011). SNVs arise as a result of errors in the DNA replication process where DNA polymerases insert incorrect nucleotides that go undetected by the genome maintenance systems, and therefore are not corrected. SNVs can also arise as a result of exposure to radiation or chemical agents. There are two types of SNVs: germline and somatic. Germline SNVs are SNVs that occur in the germ cells (cells that are destined to become the egg cell or a sperm cell) and are subsequently passed on to the offspring. Certain germline SNVs cause inherited diseases. An inherited disease is a disorder that results from a mutation in a single gene and has 100% penetrance. Examples include cystic fibrosis and sickle cell anaemia.

Somatic SNVs are SNVs that occur in somatic cells i.e. cells that are not gametes and are not passed on to offspring. Certain somatic SNVs act as drivers in cancer. DNA sequencing is used to detect germline SNVs causing inherited diseases and somatic SNVs acting as drivers in cancer (Jamuar and Tan, 2015; Chong *et al.*, 2015; Pabinger *et al.*, 2014; Watson *et al.*, 2013; Klug *et al.*, 2012).

1.2 Whole Genome Sequencing

The cost of DNA sequencing has fallen sharply in recent years beating Moore's law (Sboner *et al.*, 2011) as shown in Figure 1.1. Moore's law states that computer processing power will double every two years (Moore, 1998). Today it is possible to sequence an entire human genome for less than $1000 (Hayden, 2014). This makes it affordable for whole genome sequencing to become commonplace (see http://www.genomicsengland.co.uk/the-100000-genomes-project/). With the cost of genome sequencing continuing to fall, it will soon be cheaper to conduct whole genome sequencing instead of targeted sequencing (i.e. exome sequencing or sequencing of specific panels of genes) (Fratkin *et al.*, 2012).

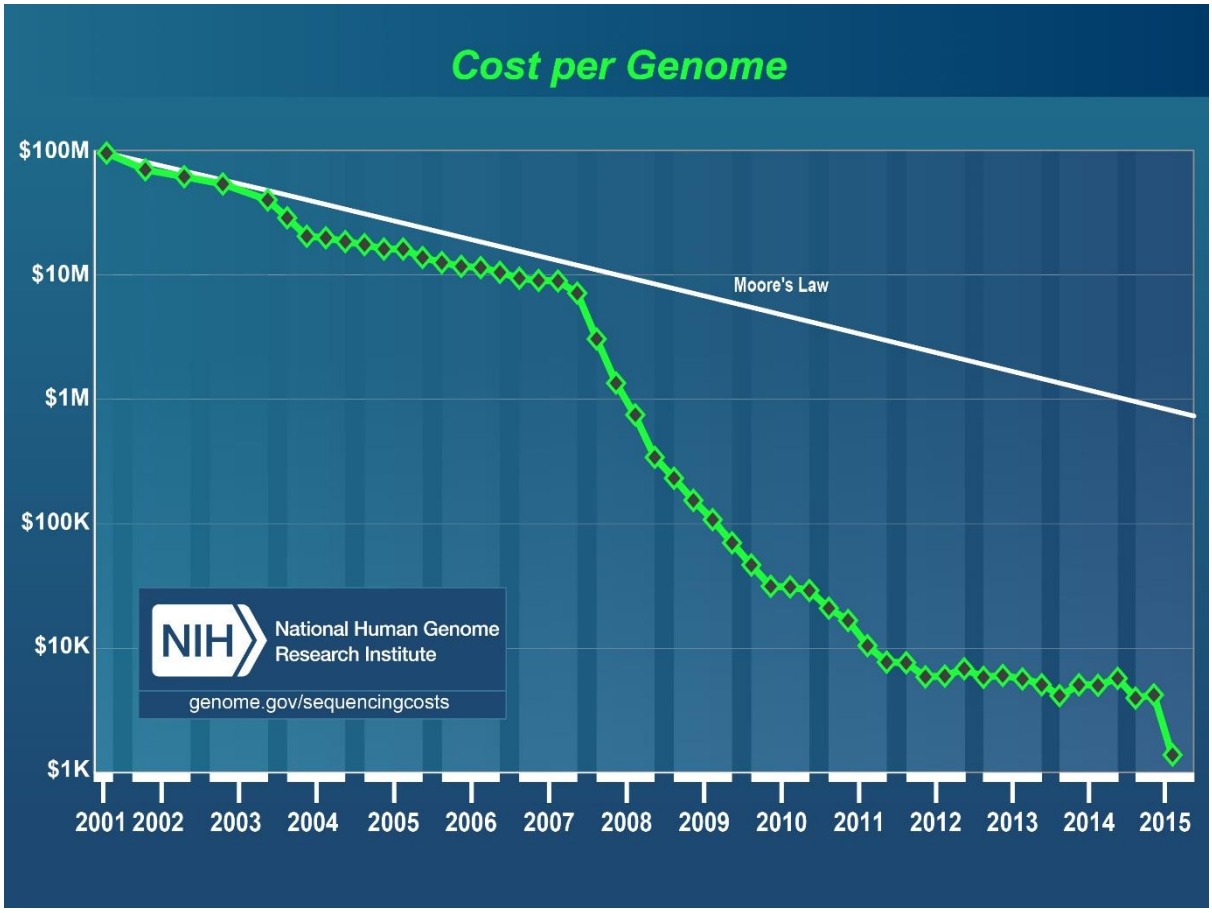


Figure 1.1: Graph showing the falling costs of genome sequencing compared with Moores Law (Reproduced from http://www.genome.gov/sequencingcosts/ last accessed on 14/05/2016).

In contrast to exome sequencing and the sequencing of specific panels of genes, the use of whole genome sequencing allows the study of both the non-coding and the coding regions in the human genome. The former comprises ~98% of the human genome, while the latter only comprises ~2% (Schnekenberg and Németh, 2013).

The non-coding region was originally described as junk DNA (Orgel and Crick, 1980). However, the recent ENCODE project showed that a large proportion of the human genome is functional (Consortium, 2012). While the

exact proportion of the genome that is functional is under debate, it is very clear that some of the non-coding region is not junk DNA, and has critical function in terms of regulating gene expression and function across cells, tissues and organs (Fratkin *et al.*, 2012; Consortium, 2012; Graur *et al.*, 2013). Therefore the non-coding region is now considered as a rich source of disease-associated SNVs that have, to date, not been properly studied, and for which the results of the ENCODE project could be of immense help in terms of developing highly precise diagnostics (Schnekenberg and Németh, 2013; Fratkin *et al.*, 2012).

1.2.1 Carrying Out Whole Genome Sequencing

In order to carry out whole genome sequencing, the DNA must first be extracted from nucleated cells in a blood or tissue sample. The extracted DNA is then randomly fragmented in order to shorten the long DNA into shorter fragments. This is a key step, as the size is extremely important for construction of the sequencing library (which is discussed in further detail below). This fragmentation is normally done by physical (i.e. sonication or acoustic shearing), or enzymatic (digestion by DNAse 1 or Fragmentase) fragmentation methods. However, enzymatic methods have been found to produce more artefacts and therefore physical methods are preferred (Marine *et al.*, 2011).

The next step is to prepare the sequencing library. This involves, first performing end repair by blunting the ends of the fragments, and then, phosphorylating the 5’ ends using the enzymes T4 Polynucleotide kinase, T4 DNA Polymerase and Klenow Large Fragment. The 3’ ends are then Poly A-

tailed (i.e. stretches of adenine nucleotides are added to the 3' ends) in order to facilitate ligation to the adaptors using the enzymes Taq polymerase, or Klenow Fragment (Adey *et al.*, 2010). Adaptors (short oligonucleotides of known sequences) are then ligated to the DNA fragments. These adaptors act as universal priming sites during the PCR amplification and sequencing stages which are discussed in more detail below. All the DNA fragments are used for cluster generation (i.e. the conversion of the sequencing library into DNA clusters) and sequencing. This is in contrast to both whole-exome sequencing and sequencing of specific gene panels, where a physical-capture step enriches the DNA fragments for the entire protein coding region in the case of whole exome sequencing, and certain genes in the case of sequencing of specific gene panels (Metzker, 2009). The absence of the capture step in whole genome sequencing results in uniform coverage, removing any areas of low coverage caused by inefficient capture. This in turn reduces the average depth of coverage that is required for accurate SNV calling, details of which are discussed in section 1.2.2.4.

Distinct clusters are then generated by spatially separating the DNA fragments, and then clonally amplifying them using PCR. DNA sequencing is then carried out. DNA can either be sequenced from one end (known as single end sequencing) or both ends (known as paired end sequencing). The latter is considered to be the better approach, as it reduces ambiguity during the alignment stage of the data analysis (Schnekenberg and Németh, 2013) (details of which are discussed further in section 1.2.2.3). There are several platforms for DNA sequencing including Illumina, Roche and SOLID. While

there are some differences in the way they operate, the basic principles are identical on all platforms as shown in Figure 1.2. These basic principles are: four nucleotides which are fluorescently labelled are first added; if the nucleotide is incorporated into the DNA, a detectable signal is generated; the signal is then converted into a base in a process known as base calling. These steps are repeated over multiple cycles resulting in thousands of DNA fragments being analysed and sequenced (Schnekenberg and Németh, 2013; Head *et al.*, 2014; Natrajan and Reis-Filho, 2011). The above steps for carrying out whole genome sequencing are summarised in Figure 1.3.

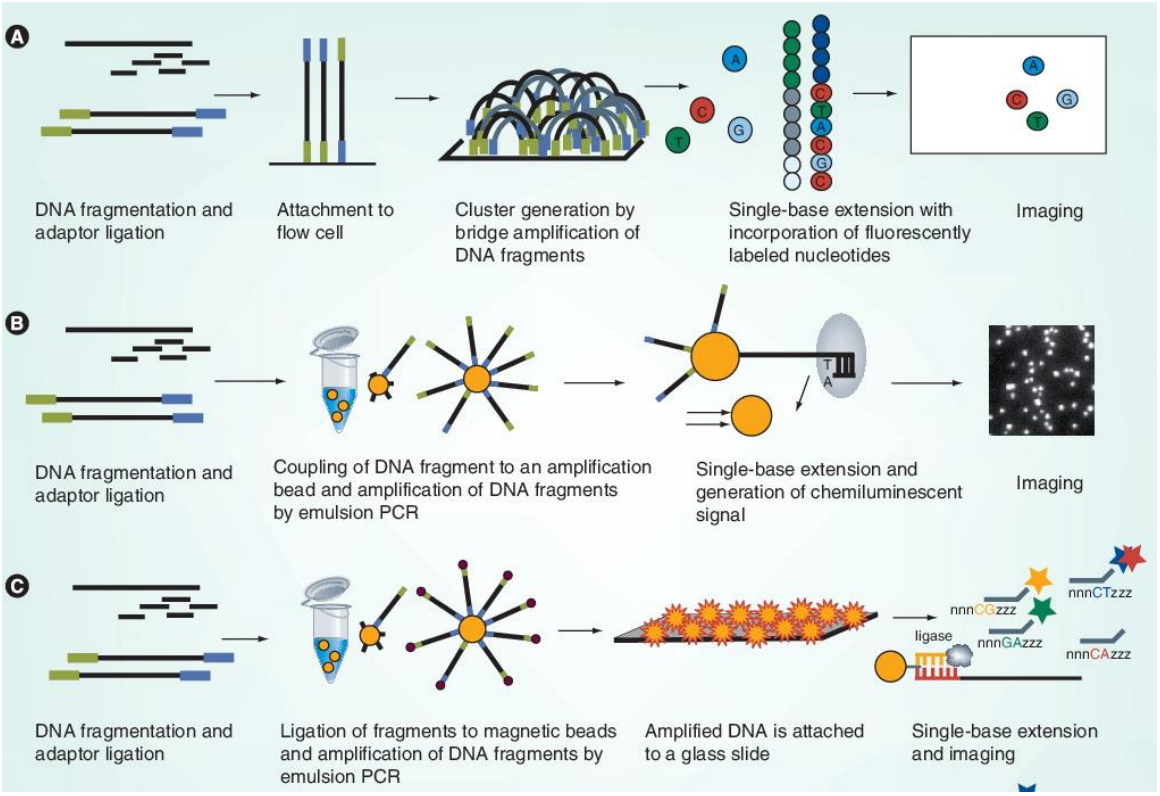


Figure 1.2: Summary of the whole genome sequencing process for Illumina (A),Roche (B) and SOLID (C) (Reproduced from (Natrajan and Reis-Filho, 2011)).

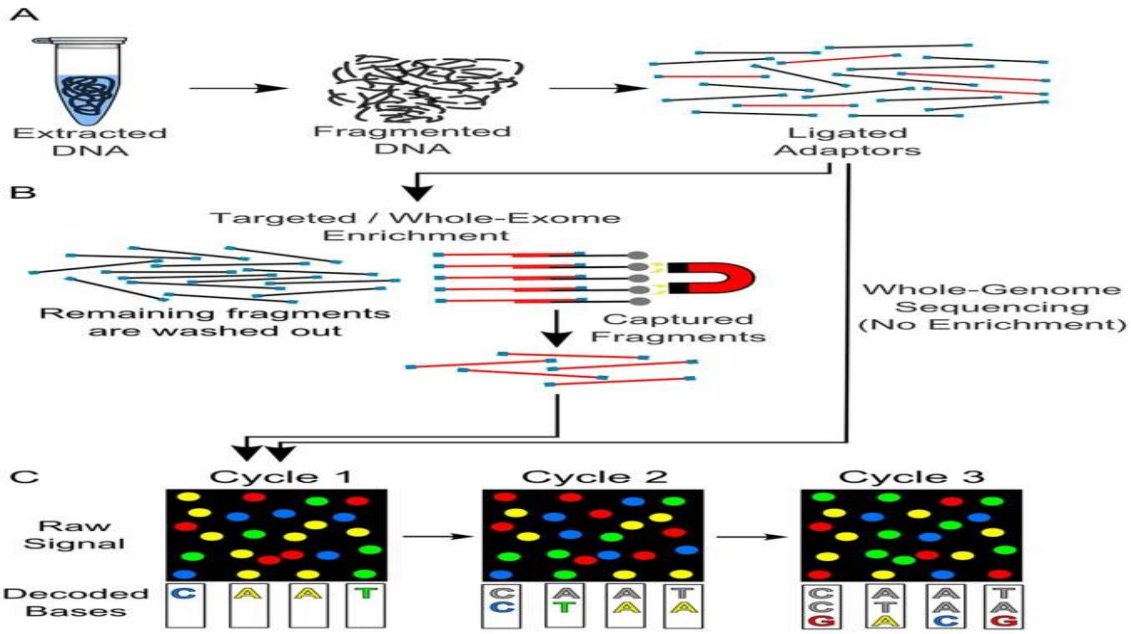


Figure 1.3: Summary of steps for carrying out whole genome sequencing (Reproduced from (Schnekenberg and Németh, 2013)).

1.2.2 Analysing Whole Genome Sequencing Data

The data obtained from whole genome sequencing consist of several million short nucleotide sequences that are about 35-400bp in length and are known as reads. These reads are provided by the sequencing platform in large (hundreds of gigabytes) text files in FASTQ format.

The FASTQ format is a plain text format for representing nucleotide sequences together with their associated quality scores. The FASTQ format consists of four records per entry. The first record begins with a '@'

character, and is followed by the sequence identifier and optionally a description. The second record consists of the sequence which can be split across multiple lines. The third record begins with a '+' character and is followed optionally by a repeat of the first record (excluding the '@' character). The fourth record consists of the quality scores and must be the same length as the second record (the sequence). These quality scores range from the '!' character (representing the lowest quality) to '~' (representing the highest quality). The fourth record can also be split across multiple lines. (Cock *et al.*, 2010). The FASTQ format is illustrated in Figure 1.4.

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#""""""""""""7F@71,'";C?,B;?6B;:EA1EA
1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@
/=<?7=9<2A8==
```

Figure 1.4: An example of the FASTQ format (Reproduced from (Cock *et al.*, 2010)).

In order to call SNVs, further analyses need to be conducted on these data which comprise several processing steps that together comprise the workflow for calling SNVs from whole genome sequencing data (Altmann *et al.*, 2012; Schnekenberg and Németh, 2013; Pabinger *et al.*, 2014). This workflow can be automated in many ways. This can involve writing scripts (either shell scripts or scripts written in a scripting language such as Perl or Python) or using tools such as Ruffus (Goodstadt, 2010), Bpipe (Sadedin *et al.*, 2012), Biopieces (http://maasha.github.io/biopieces/) ,Rake

(https://github.com/ruby/rake), SnakeMake (https://bitbucket.org/snakemake/snakemake/wiki/Home), Anduril (Ovaska *et al.*, 2010), Taverna (http://www.taverna.org.uk/) or Galaxy (Goecks *et al.*, 2010). The basic workflow is shown in Figure 1.5 and the individual steps are discussed in more detail in the following sections.

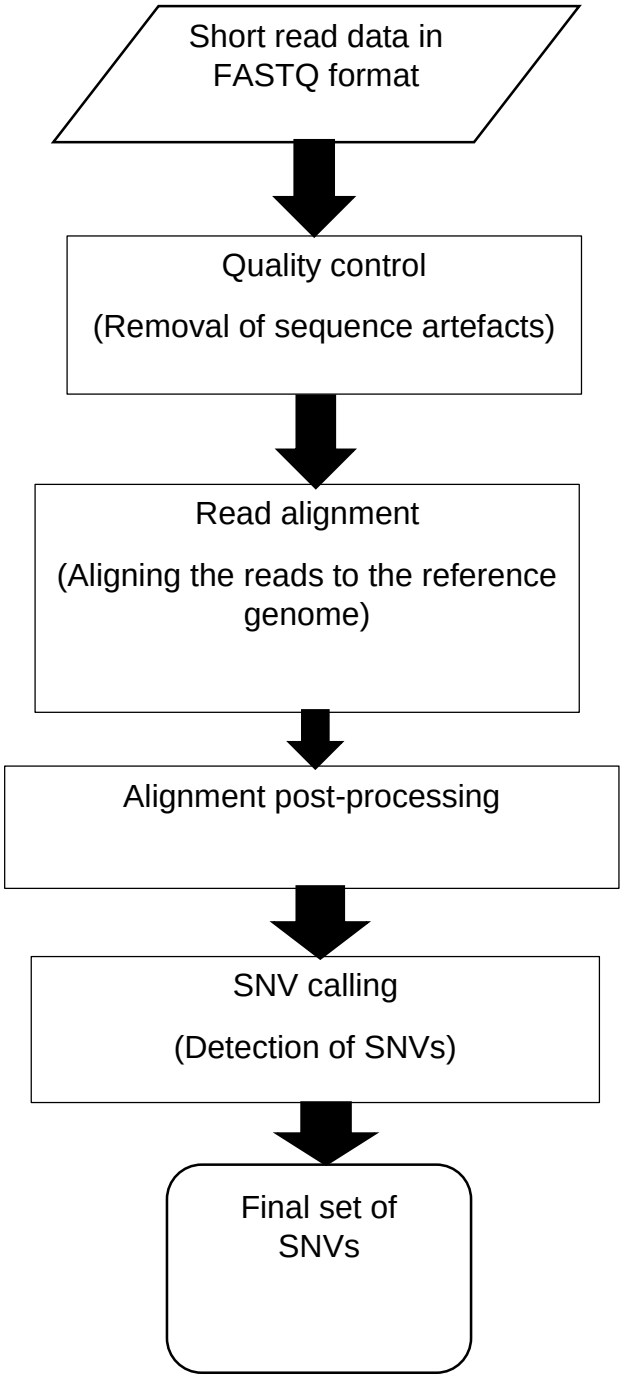


Figure 1.5: Workflow for calling SNVs from whole genome sequencing data.

The quality control and read alignment steps are common to all techniques involving next generation sequencing e.g. whole genome sequencing, RNA-

Seq, ChIP-Seq while the alignment post-processing and SNV calling steps are exclusive to genome resequencing (whole genome sequencing and exome sequencing).

1.2.2.1 Quality Control

The sequencing platforms are prone to errors in chemistry and instrumentation. Therefore the raw sequence data that are generated will contain sequence artefacts. These are errors in base calling, poor quality reads and contamination by adaptors. In order to prevent erroneous biological conclusions from being drawn as a result of these errors, it is essential to check the quality of the reads. This involves visualising the base quality scores and nucleotide distributions. Any errors are then removed by trimming the reads and/or filtering the reads based on base quality score, primer contamination and GC bias. This is done by using the FastQC package (http: //www.bioinformatics. babraham.ac.uk/projects/fastqc/) for the visualisation and using utilities in the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) to filter and trim the reads as required (Altmann *et al.*, 2012).

1.2.2.2 Read Alignment

After the reads have been processed to remove the sequence artefacts, they are aligned with the human reference genome. This involves determining the location within the human reference genome for a particular read which requires the human reference genome and an alignment tool (Pabinger *et al.*, 2014; Nielsen *et al.*, 2011).

There are currently two sources for the human reference genome: The University Of Santa Cruz (UCSC) and the Genome Reference Consortium (GRC), both of which provide multiple human genome versions. These are usually the latest version and one or more older versions. Currently UCSC provides versions hg18, hg19 and hg38 (currently the latest release) while GRC provides versions GRCh37 and GRCh38 (which is currently the latest release). The human reference genomes present in both UCSC and GRC are identical (Pabinger *et al.*, 2014).

A plethora of alignment tools have been developed over the past few years (Pabinger *et al.*, 2014; Flicek and Birney, 2009). These include Bowtie (Langmead *et al.*, 2009), Bowtie2 (Langmead and Salzberg, 2012), BWA (Li and Durbin, 2009), MAQ (Li *et al.*, 2008a), SOAP (Li *et al.*, 2008b), SOAP2 (Li *et al.*, 2009c), ZOOM (Lin *et al.*, 2008a), SHRiMP (Rumble *et al.*, 2009), BFAST (http://genome.ucla.edu/bfast/) and MOSAIK (http://bioinformatics.bc.edu/marthlab/Mosaik/). These tools follow the same fundamental procedure to align the reads to the human reference genome, and have parameters to control the number of mismatches between the reads and the human reference genome. If the parameters are set only to allow perfect matches, then it would not be possible to detect any SNVs. On the other hand, setting the parameters to allow many mismatches will result in many wrongly aligned reads and result in the calling of false positive SNVs. Therefore, the parameter settings to control the number of mismatches between the reads and the human reference genome must be carefully chosen. The procedure that the tools use to align the reads to the human reference genome exploits heuristic techniques to focus quickly on a

small set of locations in the human reference genome, where the best mapping is likely. After the identification of a smaller subset of potential mapping locations, a more accurate alignment algorithm such as Smith-Waterman is run on the smaller subset. It would be computationally infeasible to run these more accurate alignment algorithms to search all possible locations in the human genome where the reads can map. The aligned reads are stored in the sequence alignment/map (SAM) format (Flicek and Birney, 2009; Altmann *et al.*, 2012), a plain text format for storing read alignments. All lines are tab delimited. The SAM format consists of two sections: a header section, and an alignment section with each line in the header section beginning with a '@' character. Each line in the alignment section has eleven compulsory fields. The compulsory fields are:

1. the read name
2. a bitwise flag providing extra information about the read
3. the reference sequence name
4. the chromosome name
5. the position of the first matching base
6. the mapping quality score
7. a string describing the pairwise alignment (this string reports the number of mismatches ('M'), the number of insertions ('I'), the number of deletions ('D'), the number of skipped bases ('N'), the number of bases not in the alignment which have been retained in the sequence ('S'), the number of bases not in the alignment which have been

excluded from the sequence ('H') and if the read has been fully aligned ('P'))

8. the chromosome name of the next read (which is reported with an '=' character if it is the same chromosome as the previous read)
9. position of the next read, the inferred insert size (approximate size of any insertions and deletions with deletions reported as negative numbers)
10. the sequence
11. The base quality score

Any unavailable information is represented with a '*' character or a zero (Li *et al.*, 2009b). The SAM format is illustrated in Figure 1.6.

```
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTA *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA    *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC     *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

Figure 1.6: An example of the SAM format (Reproduced from (Li *et al.*, 2009b)).

There are two types of heuristic alignment techniques: hash-based and the Burrows Wheeler Transform (BWT). Hash-based methods are based on the use of a hash table data structure to scan and index the sequence data. The hash table is able to allow rapid access to information on the location of subsequences within the reference genome. The hash table can be built either on the set of input reads, or on the human reference genome. In the

case of the hash table being built on the set of input reads, the reference genome is used to scan the hash table of reads, whereas, in the case of the hash table being built on the human reference genome, the set of input reads is used to scan the hash table of the reference genome. Tools that build hash tables on the reference genome have a constant memory requirement. However, this requirement depends on the size and complexity of the reference (and will be large in the case of the human reference genome), while tools that build hash tables on the set of input reads have memory requirements that are smaller and more variable, but the processing time to scan the entire genome can be greater if there are fewer input reads (Flicek and Birney, 2009). Examples of tools that utilise the hash-based method by building a hash table of the input reads are: MAQ (Li *et al.*, 2008a), ZOOM (Lin *et al.*, 2008a) and SHRiMP (Rumble *et al.*, 2009), while tools that build a hash table of the reference genome are: SOAP (Li *et al.*, 2008b), BFAST (http://genome.ucla.edu/bfast/) and MOSAIK (http://bioinformatics.bc.edu/marthlab/Mosaik/).

Methods that make use of BWT (Burrows and Wheeler, 1994) exploit the 'FM index' data structure (Ferragina and Manzini, 2000) which enables rapid sub-sequence search, and, in the case of the human (and indeed all other mammalian genomes), is equal to, or smaller in size than, the reference genome itself. There are two steps involved in creating the FM index. Firstly, BWT is used for efficient data compression of the reference genome. Secondly, the final index is created. This step can be memory intensive, but can be done in less memory with a cost in processing time. This final index

is then used for rapid placement of reads on the human reference genome. Methods utilising BWT have a greater processing speed than methods utilising hash tables (Flicek and Birney, 2009; Altmann *et al.*, 2012; Kärkkäinen, 2007). Examples of tools that utilise BWT are Bowtie (Langmead *et al.*, 2009), Bowtie2 (Langmead and Salzberg, 2012), BWA (Li and Durbin, 2009) and SOAP2 (Li *et al.*, 2009c).

1.2.2.3 Alignment Post-Processing

Once the reads have been aligned to the reference genome, several post-processing steps need to be carried out on the aligned reads before variant calling is carried out (Altmann *et al.*, 2012). First, the proportion of the reads that were successfully aligned needs to be obtained. Next, the aligned reads need to be sorted according to their position in the chromosome.

Since the PCR that is used for amplification and ligation of adaptors can introduce duplicated reads, these need to be removed. Some reads can have more than one optimal alignment: 'non-unique alignments'; which also need to be removed (Altmann *et al.*, 2012; Pabinger *et al.*, 2014). The above post-processing steps are carried out using either the SAMtools (Li *et al.*, 2009b) or Picard (http://broadinstitute.github.io/picard/) suites of tools for manipulating aligned reads (Altmann *et al.*, 2012).

Finally, reads around small indels need to be realigned to prevent calling of false-positive SNVs (Altmann *et al.*, 2012). This post-processing step is performed by utilities in the GATK suite of tools which are a set of tools for analysing next-generation sequencing data (McKenna *et al.*, 2010).

1.2.2.4 SNV Calling

After post-processing of the aligned reads, the next step is to 'call' SNVs.

There are several tools available to do this from whole genome sequencing data. These include SAMtools (Li *et al.*, 2009b), GATK (McKenna *et al.*, 2010), VarScan 2 (Koboldt *et al.*, 2012), SNVer (Wei *et al.*, 2011) and SomaticSniper (Larson *et al.*, 2012). Some tools can call both germline and somatic SNVs (e.g. SAMtools and VarScan 2) while others can only call germline SNVs (e.g. GATK and SNVer) or somatic SNVs (e.g. SomaticSniper).

The SNV calling tools all generate output in VCF format, a plain text format for storing SNV data. It consists of two sections: a header section and a data section. The header section consists of a number of meta-information lines with each one prefixed by the characters '##'. These meta-information lines describe the tags and annotations used in the data section as well as file creation information, reference genome version, software used to call SNVs and any other relevant information. The header section also contains a tab delimited field definition line which is prefixed by the character '#'. The field definition line names the eight compulsory fields. These are:

1. the chromosome (CHROM)
2. the position (POS)
3. the unique identifier (ID)
4. the reference allele (REF)
5. the mutation (ALT)
6. the quality score (QUAL)
7. filtering information (FILTER)
8. annotations (INFO)

The data section contains the data that correspond to the above fields. The lines in the data section are tab delimited and must match the number of fields defined in the header section (Danecek *et al.*, 2011). The VCF format is illustrated in Figure 1.7.

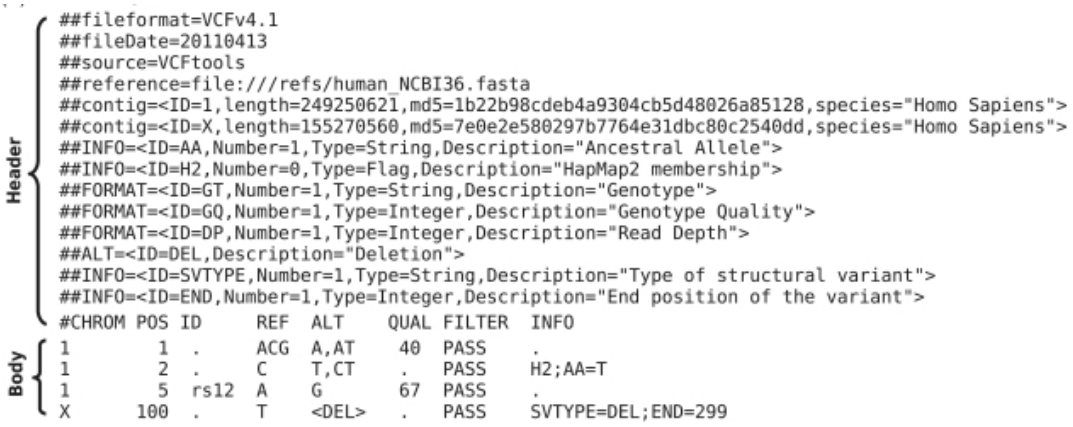


Figure 1.7: An example of the VCF format (Reproduced from (Danecek *et al.*, 2011)).

1.3 Identifying the Functional Consequence of SNVs

Once a set of SNVs have been called from whole genome sequence data the next step is to identify the functional consequence of the SNV. This in turn will enable the identification of SNVs that cause Mendelian diseases and act as driver SNVs in cancer. The possible functional consequences are: non-synonymous, synonymous, nonsense, sense, splice site and transcription factor binding site (Makrythanasis and Antonarakis, 2011). Non-synonymous, synonymous, nonsense and sense SNVs occur only in the coding region. Splice site SNVs occur in the intron-exon boundary, while transcription factor binding site SNVs occur in both the coding and non-coding regions (in promoters, introns, exons and regions far upstream of genes (up to 10,000 bp)).

At the whole genome scale, it is not feasible to employ experimental methods to identify the functional consequence of SNVs. Therefore computational approaches are required to identify the functional consequence of SNVs. This is done by the following tools ANNOVAR (Wang *et al.*, 2010), Ensembl VEP (McLaren *et al.*, 2010) and snpEff (Cingolani *et al.*, 2012).

SNVs that have different functional consequences differ in terms of their impact on the resulting protein product. Non-synonymous SNVs (also known as missense SNVs) are SNVs where one codon is replaced with another that encodes a different amino acid. (Read and Donnai, 2011; Khan and Vihinen, 2007). This is illustrated in Figure 1.8.

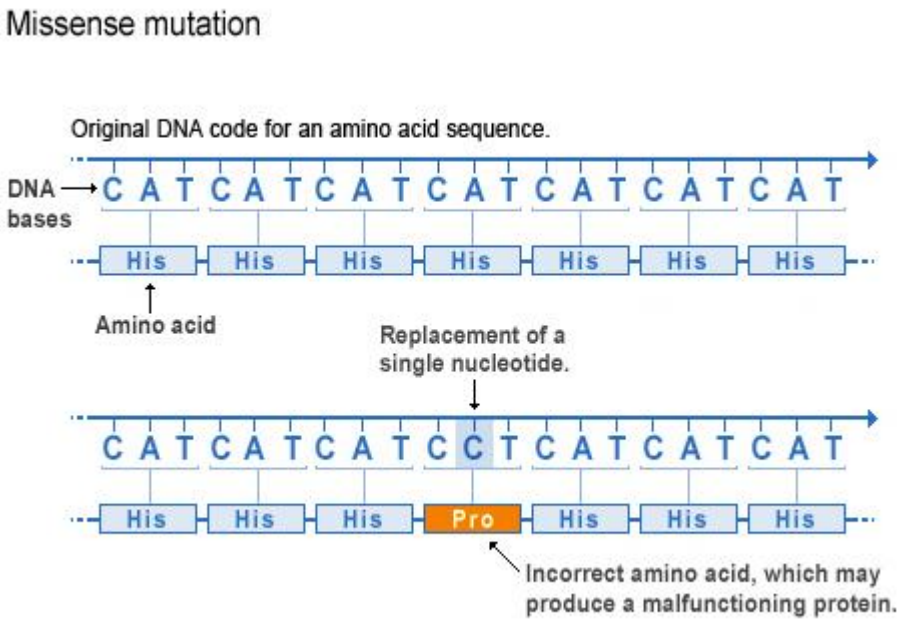


Figure 1.8: Effect of a non-synonymous SNV. (Reproduced from (http://ghr.nlm.nih.gov/handbook/illustrations/missense) last accessed on 14/05/2016).

The impact of non-synonymous SNVs on the protein is varied. Certain non-synonymous SNVs can severely damage the protein product, and alter its normal function, while others have a negligible effect on the normal function of the protein. Therefore, in addition to identifying the functional consequence of an SNV as being non-synonymous, its impact on the protein will also need to be quantified.

A plethora of tools have been developed to predict the impact of a non-synonymous SNV on protein structure and function (Khan and Vihinen, 2007; Cline and Karchin, 2011; Pabinger *et al.*, 2014). These include SIFT (Kumar *et al.*, 2009), PolyPhen-2 (Adzhubei *et al.*, 2010), FATHMM (Shihab *et al.*, 2013), MutationAssessor (Reva *et al.*, 2011), SNPs3D (Yue *et al.*,

2006), nsSNPAnalyzer (Bao *et al.*, 2005), SNAP (Bromberg and Rost, 2007), SAAPpred (Al-Numair and Martin, 2013), MutPred (Li *et al.*, 2009a), SNPS&GO (Calabrese *et al.*, 2009) , SNPs&GO3D (Capriotti and Altman, 2011), Panther (Thomas *et al.*, 2003), PhD-SNP (Capriotti *et al.*, 2006), PMut (Ferrer-Costa *et al.*, 2004), MAPP (Stone and Sidow, 2005), SusPect (Yates *et al.*, 2014), Bongo (Cheng *et al.*, 2008), Hansa (Acharya and Nagarajaram, 2012), Parepro (Tian *et al.*, 2007), SNPDryad (Wong and Zhang, 2014), Condel (Gonzalez-Perez and Lopez-Bigas, 2011) and CAROL (Lopes *et al.*, 2012).

Several of these tools make use of only sequence information to predict the impact of non-synonymous SNVs (e.g. SIFT, FATHMM, MutationAssessor, Panther, PhD-SNP, MAPP, SNPS&GO, Parepro and SNPDryad), while others make use of both sequence and structural information (e.g. PolyPhen-2, SNPs3D, nsSNPAnalyzer, SNAP, MutPred, SNPs&GO3D, PMut, SusPect, SAAPpred and Hansa). On the other hand, certain tools exclusively make use of structural information (e.g. Bongo). Recently, tools such as Condel and CAROL have been developed that exploit the complementarity of different tools for predicting the impact of non-synonymous SNVs on protein function. These tools first obtain the output from several tools, and then combine the output scores of these tools and give the final prediction of the impact of the non-synonymous SNV.

Synonymous SNVs (also known as silent SNVs) are SNVs where one codon is replaced with another that encodes the same amino acid. Synonymous SNVs are conventionally assumed to have no effect on the protein and thus

be neutral (Read and Donnai, 2011). Therefore, no further analysis needs to be done after identifying an SNV as being synonymous.

Nonsense SNVs are SNVs where a codon is replaced with a stop codon. Protein synthesis then stops at that point as illustrated in Figure 1.9. If the SNV occurs at a location which has an exon-exon junction more than 50-55 nucleotides upstream then nonsense-mediated decay ensues. This results in degradation of the mRNA transcript and complete lack of production of the protein product. This is equivalent to deletion of the entire gene. If a nonsense SNV does not trigger nonsense-mediated decay, then a truncated protein is produced as illustrated in Figure 1.10. Either way, nonsense SNVs are considered as loss-of-function SNVs (Read and Donnai, 2011; Maquat, 2005; Kurmangaliyev *et al.*, 2013). Therefore, no further analysis needs to be done after identifying an SNV as being nonsense.

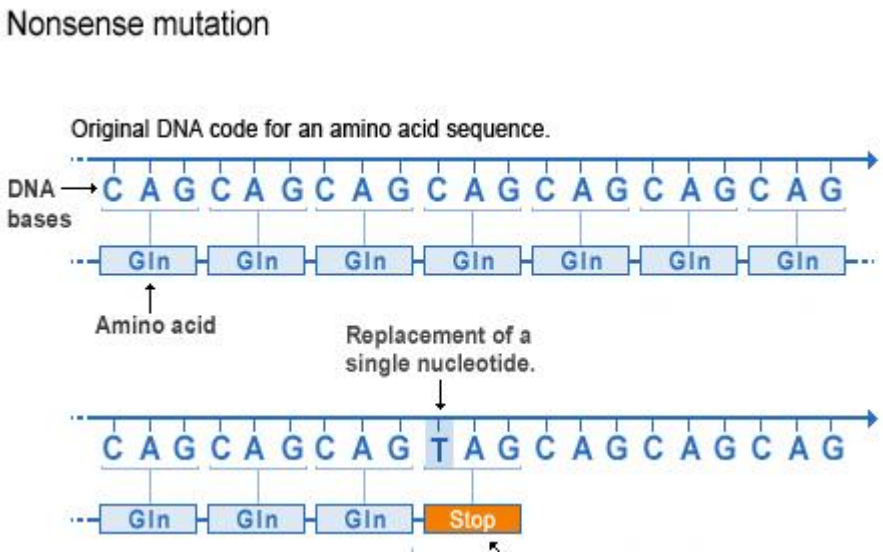


Figure 1.9: Effect of a nonsense SNV (Reproduced from (http://ghr.nlm.nih.gov/handbook/illustrations/nonsense) last accessed on 14/05/2016).

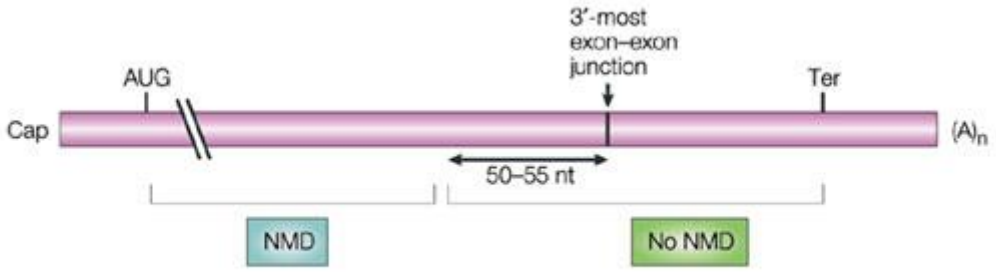


Figure 1.10: Locations of the gene where nonsense SNVs trigger and do not trigger nonsense mediated decay (Reproduced from (Maquat, 2004)).

Sense SNVs are SNVs where a stop codon is replaced with a codon that codes for an amino acid. This results in the downstream 3' Untranslated Region becoming part of the open reading frame which will result in a protein with a C-terminal extension. The mRNA transcript is hence degraded as the resulting protein product will be unstable. Therefore there is a complete lack

of production of the protein product which is equivalent to complete gene deletion (Klauer and van Hoof, 2012). Therefore no further analysis needs to be done after identifying an SNV as being sense.

SNVs that occur in splice sites disrupt the splice sites located in the intron-exon boundary that are required for the removal of introns and the joining of the exons which in turn yields the mature mRNA molecule. This results in the skipping of the relevant exon, or retention of intronic sequence therefore yielding a non-functional copy of the protein product as shown in Figure 1.11 . Splice site SNVs are also considered as loss of function SNVs (Read and Donnai, 2011; Kurmangaliyev *et al.*, 2013). Therefore no further analysis needs to be done after identifying an SNV as being in a splice site.

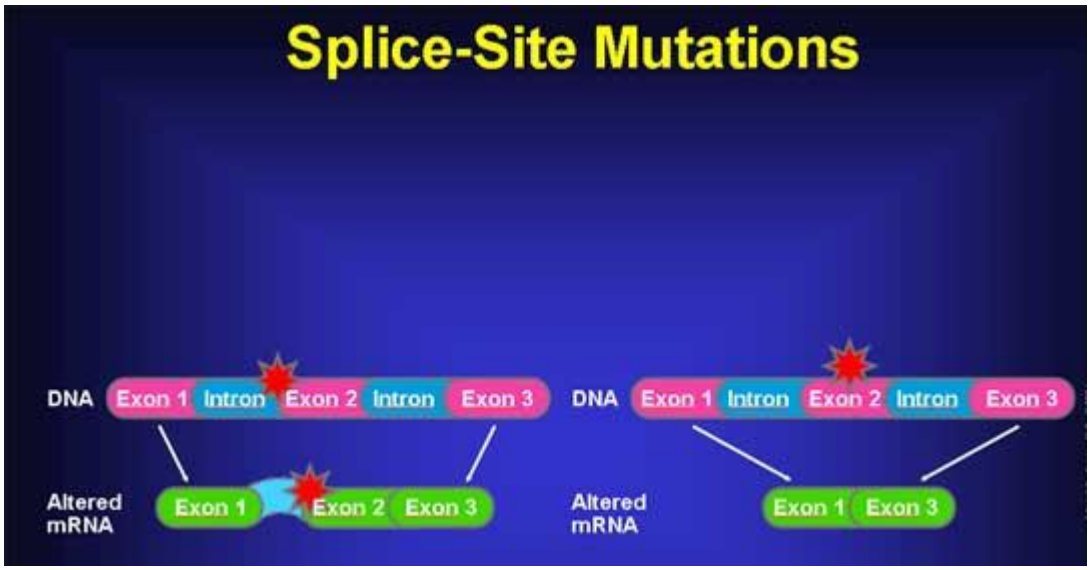


Figure 1.11: Effect of a splice site SNV (Reproduced from (http://www.cancer.gov/) last accessed on 14/05/2016).

SNVs that occur in TFBSs can disrupt the key protein-DNA interactions required for binding of transcription factors to their corresponding TFBSs which regulate transcription as shown in Figure 1.12. Gene expression of the corresponding gene is therefore altered. Consequently the mRNA and hence protein levels, are altered (Worsley-Hunt *et al.*, 2011; de Vooght *et al.*, 2009). However, this is now thought only to be the consequence of SNVs occurring in TFBSs in the non-coding regions as questions have been raised as to whether TFBSs in protein coding regions are functional in terms of regulation of gene expression (Xing and He, 2015). In order to identify SNVs that occur in TFBSs, a set of TFBSs are required.

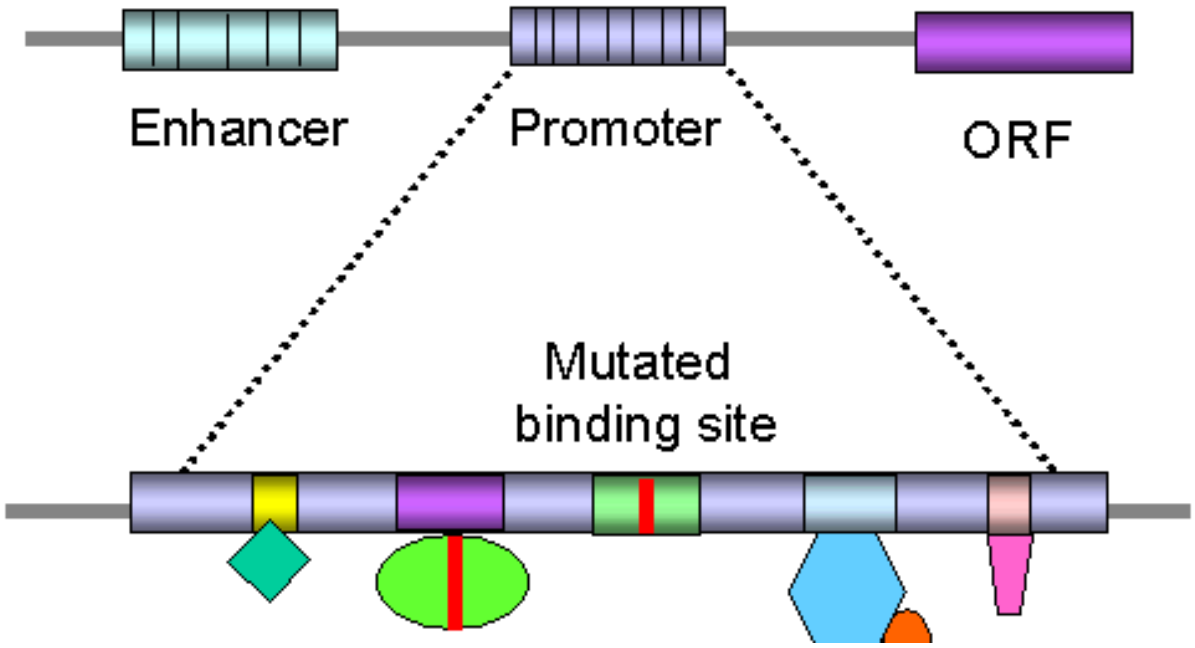


Figure 1.12: Effect of an SNV in a transcription factor binding site (Reproduced from (http://www.gene-regulation.com/info/pathodb.html) last accessed on 14/05/2016).

1.4 Experimental Identification of Transcription Factor Binding Sites

There are many experimental techniques that have been used to identify TFBSs. These are reviewed briefly below.

Traditionally, the Electro-Mobility Shift Assay (EMSA) (Garner and Revzin, 1981) has been the *de facto* technique for experimentally identifying TFBSs. EMSA is carried out by subjecting mixtures of protein and nucleic acid to electrophoresis, and then using autoradiography to determine the distribution of nucleic acid mixtures. EMSA works by exploiting the ability of a non-

denaturing polyacrylamide gel to act as a molecular sieve, hence separating the protein-bound DNA from the unbound DNA. The protein-nucleic acid complexes migrate more slowly than the free nucleic acid (Hellman and Fried, 2007; Elnitski *et al.*, 2006). The EMSA assay is summarised in Figure 1.13.

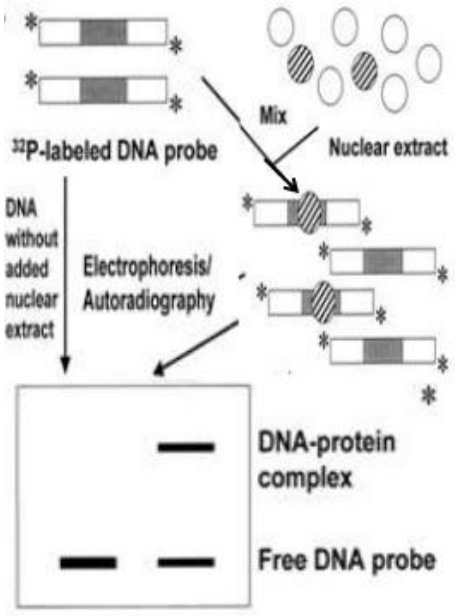


Figure 1.13: The EMSA assay (Reproduced from (Yang, 1998)). Electrophoresis is carried out on mixtures of protein and ^{32}P labelled nucleic acid. This is followed by autoradiography. The unbound DNA separates from the protein-nucleic acid complexes by migrating faster on the gel.

An alternative technique is the DNase I footprinting/protection assay which combines the cleavage reaction of DNase I with the binding properties of the EMSA assay (Galas and Schmitz, 1978). The fundamental principle of the DNase I footprinting/protection assay is that the bound protein protects the phosphodiester backbone of the DNA from hydrolysis by DNase I. Following hydrolysis by DNase I, the resulting fragments undergo electrophoresis and are visualised by autoradiography. Any TFBSs that are cleavage-protected will appear as a blank image in the semicontinuous ladder of nucleotide positions (Brenowitz *et al.*, 2001; Elnitski *et al.*, 2006). The DNase I footprinting/protection assay is summarised in Figure 1.14.

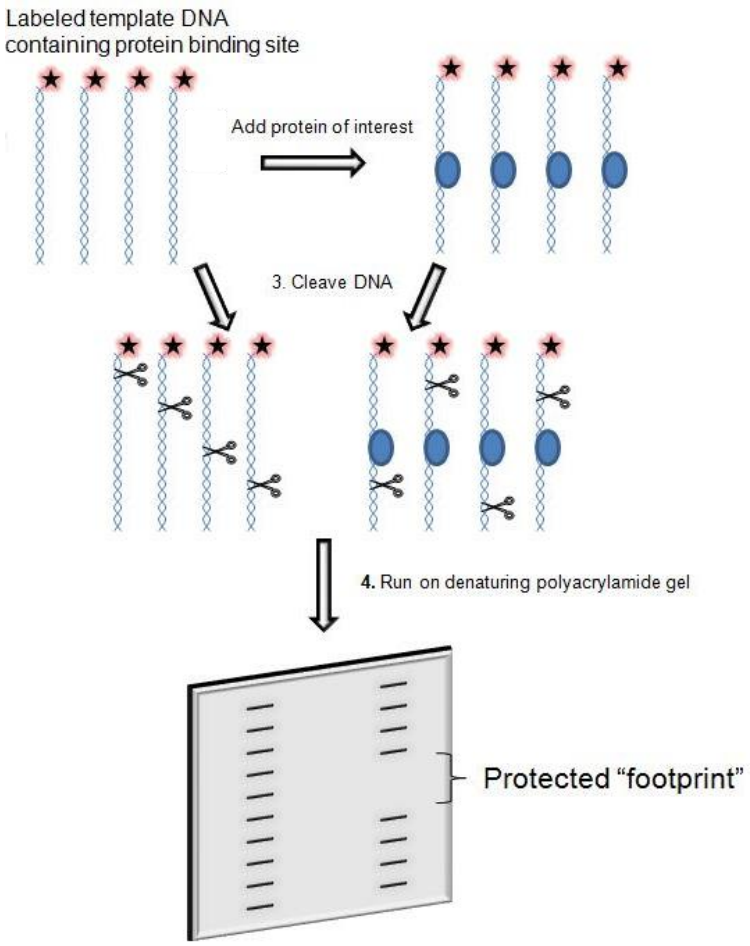


Figure 1.14: The DNase I footprinting/protection assay. Reproduced from (www.nationaldiagnostics.com) last accessed on 14/05/2016). The labelled DNA is hydrolysed by DNase I and the resulting fragments undergo electrophoresis and are visualised by autoradiography. Areas of the DNA bound by protein are protected from hydrolysis and appear as blank images.

A key problem with both the EMSA and DNase I footprinting/protection assays is the identification of unwanted protein-DNA interactions that result from the interference of non-specific DNA binding proteins such as DNA repair proteins (Elnitski *et al.*, 2006).

A more technically advanced assay is the ‘Systematic Evolution of Ligands by EXponential enrichment’ (SELEX) assay (Tuerk and Gold, 1990). SELEX is used to select dsDNAs that are bound specifically by a particular transcription factor from a random library. It works by screening a large pool of short, random oligonucleotide probes which are recognized by a protein of interest (Tuerk and Gold, 1990). The oligonucleotides that are bound by the protein of interest are then separated from the oligonucleotides that are not bound in a step known as selection. The oligonucleotides that were bound by the target protein are then amplified by PCR. This process of screening, selection and PCR amplification is termed a SELEX ‘round’. Multiple rounds of SELEX are performed (Tuerk and Gold, 1990; Djordjevic, 2007). The SELEX protocol is summarised in Figure 1.15.

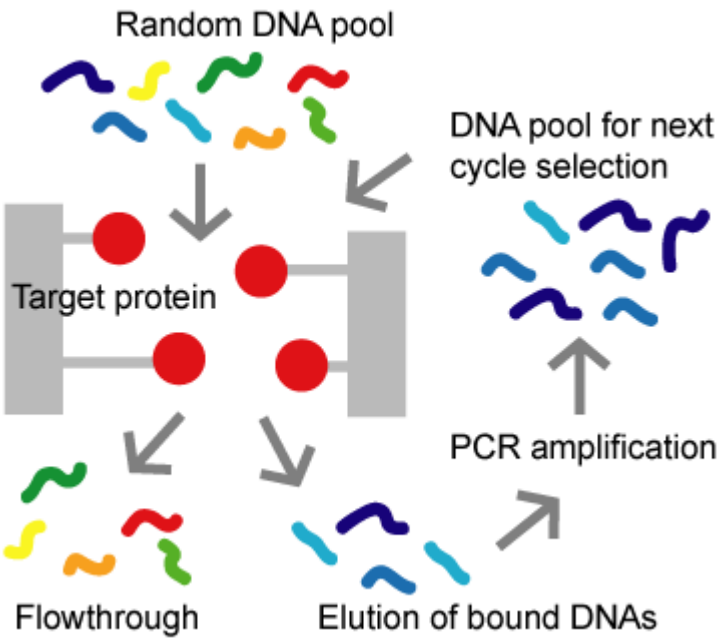


Figure 1.15: The SELEX assay Reproduced from (altair.sci.hokudai.ac.jp) last accessed on 14/05/2016). SELEX consists of multiple rounds. A SELEX round consists of screening a large pool of short nucleotide probes that are recognised by the protein of interest. Oligonucleotides that are bound to protein are separated from oligonucleotides that are free. PCR amplification of the protein-bound oligonucleotides then follows.

1.5 Experimental Identification of Genome Wide Transcription Factor Binding Events

There is now an opportunity to identify and characterize protein-DNA binding events at a genome-wide level through the use of the techniques ChIP-chip and ChIP-Seq. ChIP-chip and ChIP-Seq are high throughput versions of the Chromatin ImmunoPrecipitation (ChIP) assay.

In a ChIP assay, the DNA-binding protein of interest is cross-linked to the DNA using formaldehyde, hence capturing the protein-DNA interactions *in vivo*. The DNA is then fragmented into small fragments of around 200–1000 bp, and an antibody specific for a given transcription factor is then used to immunoprecipitate the DNA-protein complex. The cross-links are then reversed, releasing the DNA for PCR amplification (Elnitski *et al.*, 2006). However, the ChIP assay has the inability to detect the precise binding sites (that are between 9 and 15 bp long) within the identified regions. The ChIP assay is summarised in Figure 1.16.

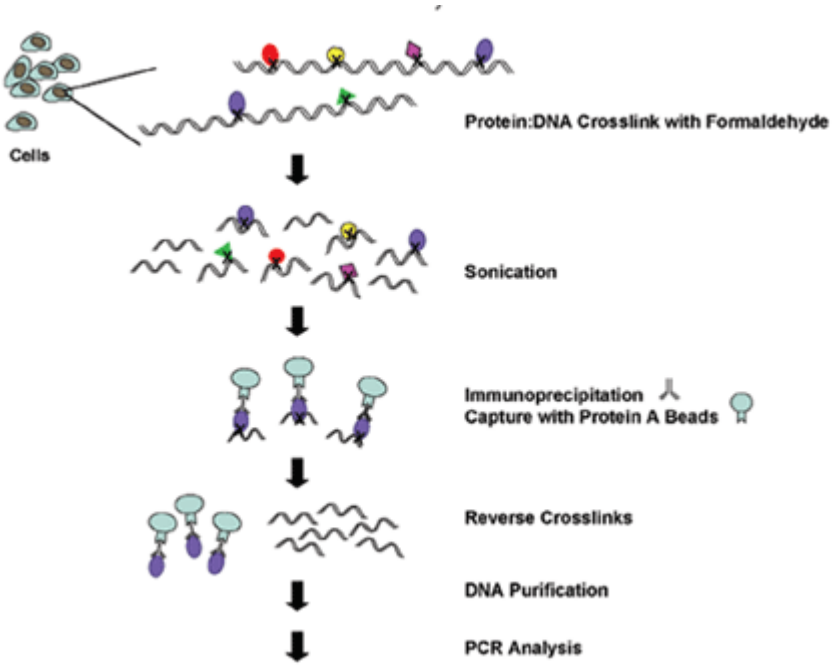


Figure 1.16: The ChIP assay (Reproduced from (www.seoulin.co.kr) last accessed on 14/05/2016). Formaldehyde is used to cross-link a DNA binding protein to the DNA. The DNA is then fragmented and the DNA-protein complex is immunoprecipitated. The cross-links are reversed and the DNA undergoes PCR amplification.

ChIP-chip involves labelling the resulting fragments from the ChIP assay with a fluorescent molecule (e.g. Cy5 or Alexa 647) followed by hybridization to genomic tiling microarrays (Ren *et al.*, 2000). The labelling and hybridization steps are similar to cDNA microarrays (Ren *et al.*, 2000; Buck and Lieb, 2004). The ChIP-chip experiment is summarised in Figure 1.17. ChIP-Seq involves performing end repair, poly-A tailing and then the ligation of adaptors to the resulting DNA fragments from the ChIP assay. Clusters of

these fragments are then generated. Massively parallel sequencing is then carried out (Park, 2009). These steps are similar to other next generation sequencing experiments such as whole genome or RNA sequencing. The ChIP-Seq experiment is summarised in Figure 1.18.

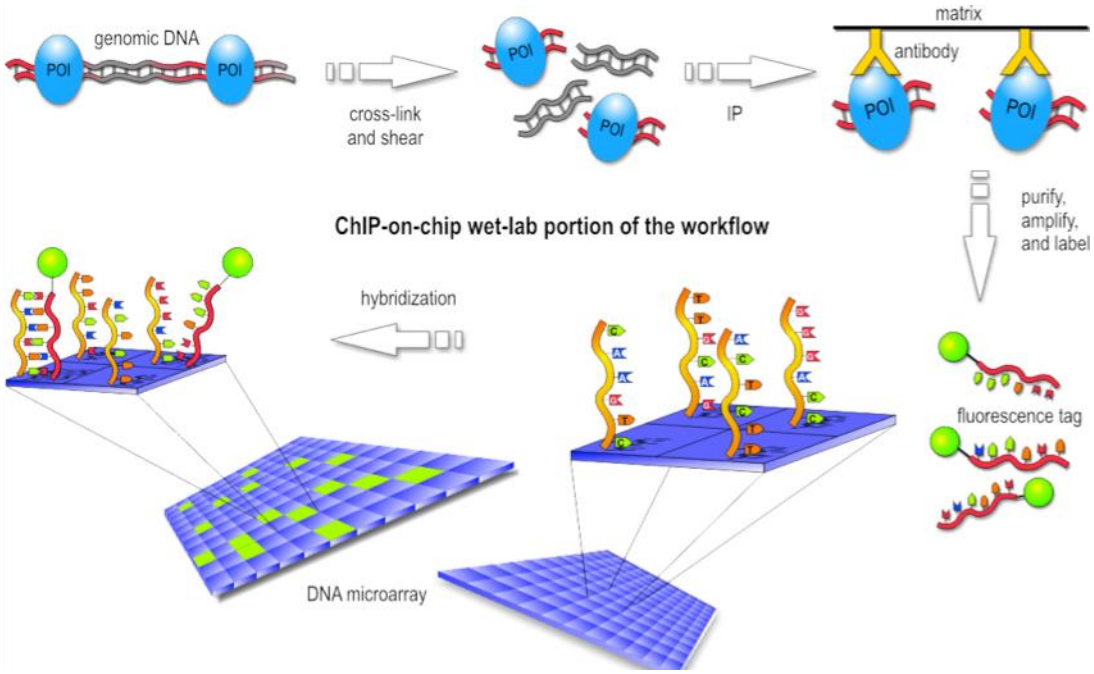


Figure 1.17: The ChIP-ChIP workflow (Reproduced from www.bcm.edu last accessed on 14/05/2016). The resulting fragments from the ChIP assay are labelled with a fluorescent molecule and then hybridised to a microarray.

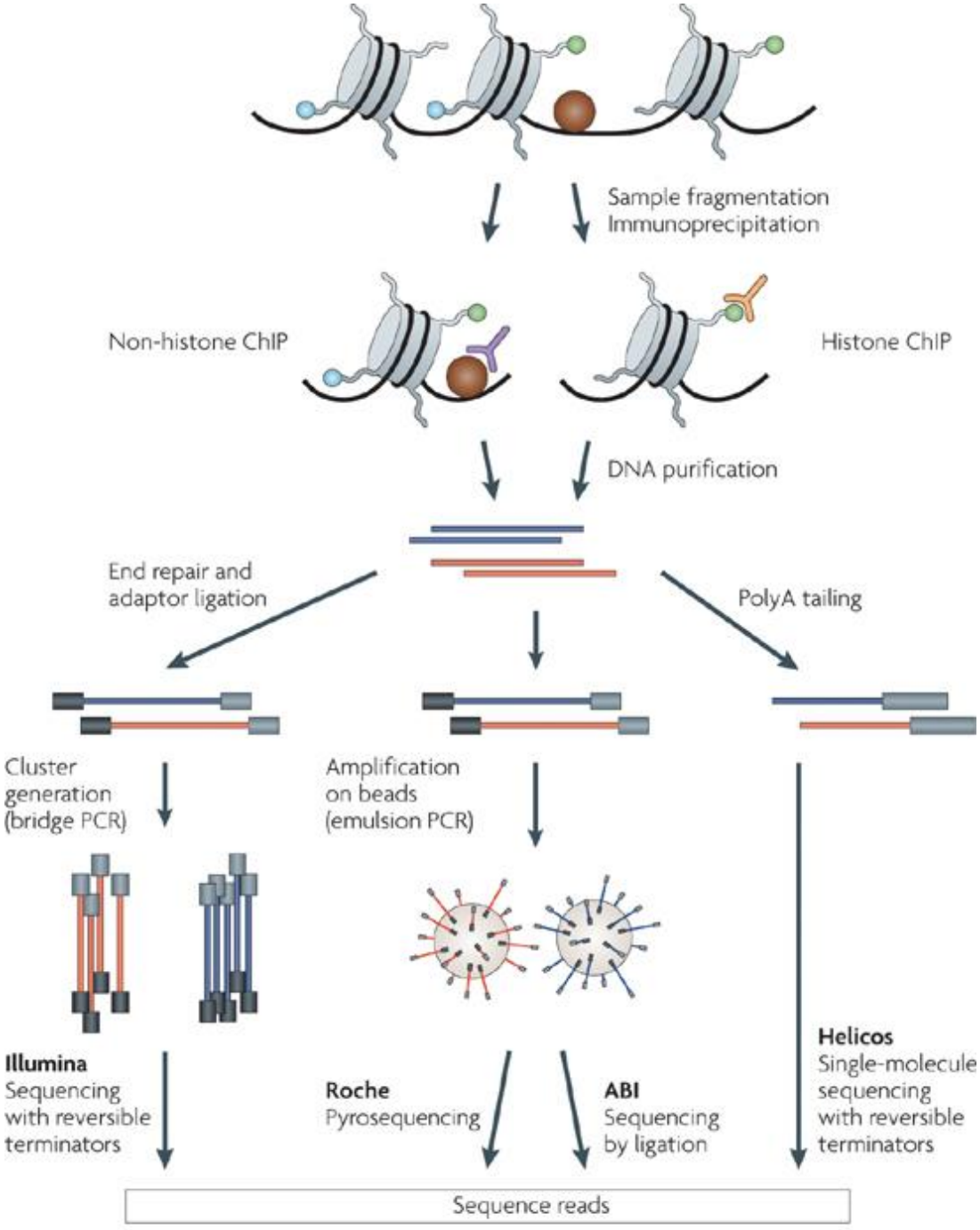


Figure 1.18: The ChIP-Seq Workflow (Reproduced from (Park, 2009)). End repair, poly-A tailing and adaptor ligation is carried out on the DNA fragments resulting from the ChIP assay. Clusters of fragments are then generated which are then subject to massively parallel sequencing.

There are a number of advantages of using ChIP-Seq instead of ChIP-chip to identify transcription factor binding regions. A key improvement over ChIP-chip is in base pair resolution given that arrays have limitations in resolution arising from uncertainties in the hybridization step which can also introduce noise from cross hybridization between sequences that are not perfectly matched. This arises owing to the inherent complexity of nucleic acid hybridization, and the fact that it depends on multiple factors such as GC content, length, concentration, secondary structure of the target and probe sequences (Park, 2009). In addition, the intensity signal measured on arrays suffers from non-linearity over its range (Park, 2009) and the dynamic range can also be limited such that the signal is below the sensitivity threshold or above the saturation point. The result is that biologically relevant peaks which are identified in ChIP-Seq are obscured when ChIP-chip is employed. In addition, ChIP-Seq allows repetitive regions to be analysed which are normally obscured on arrays. This is facilitated by the fact that genomic coverage is not limited to the probe sequences that have been fixed on the array in the ChIP-chip approach. Hence ChIP-Seq has a higher specificity and sensitivity compared with ChIP-chip (Park, 2009; Joshua *et al.*, 2011), and has largely superseded the ChIP-chip method. ChIP-Seq is now the current gold standard for identifying protein/DNA interaction regions such as transcription factor binding regions (Adli and Bernstein, 2011).

1.6 Challenges in Identifying SNVs in Transcription Factor Binding Sites

Unfortunately the number of precise experimentally characterised TFBSs is very limited. This is because the techniques that experimentally identify precise TFBSs (EMSA, DNase I footprinting/protection and SELEX assays) are low-throughput and hence are only able to characterize a small number of protein-DNA binding events. However, the recent ENCODE project has resulted in a large amount of ChIP-Seq data being publicly available. Therefore, for many transcription factors, whole genome maps of transcription factor binding exist. However, as the binding regions identified by these ChIP-Seq experiments are much longer than the binding site for a particular transcription factor, the precise TFBS within a region identified by ChIP-Seq still needs to be detected. Given that the number of experimentally characterised TFBSs are limited, the detection of the precise TFBS within a ChIP-Seq region is completely reliant on computational prediction of TFBSs. The computational prediction of TFBSs is generally performed by using a pattern matching tool to scan ChIP-Seq regions with a Position Weight Matrix (PWM) which describes a transcription factor of interest. Improving the computational prediction of TFBSs will aid the identification and interpretation of SNVs in TFBSs from whole genome sequencing data (Worsley-Hunt *et al.*, 2011; Consortium, 2012; Fratkin *et al.*, 2012; Hunt *et al.*, 2014; Bailey and Machanick, 2012).

1.7 Aims and Outline of Thesis

This thesis will focus on improving the computational prediction of TFBSs, and, then exploiting the resulting predicted TFBSs to interpret the effects of SNVs in TFBSs in particular focussing on driver and passenger SNVs in cancer.

Chapter 2 discusses the improvement of the computational prediction of TFBSs through the evaluation of the performance of a set of pattern matching tools that can be locally installed, and the identification of the best performing tool. Chapter 3 discusses the improvement of the computational prediction of TFBSs through the evaluation of a set of motif discovery tools that are used to derive PWMs, the identification of the best performing tool, the use of this tool to generate a set of new PWMs and by finally checking that the selection of the best pattern matching tool is not unduly influenced by the choice of PWMs. Chapter 4 discusses the application of the analyses in chapters 2 and 3 to the problem of discriminating between somatic cancer driver and passenger SNVs in TFBSs. Chapter 5 provides a summary of the major findings in this thesis and discusses future work.

2 An Independent Assessment of Pattern Matching Tools

The work presented in this chapter was presented as a poster at the MASAMB 2013 conference held at Imperial College London and has been submitted as a paper to BMC Bioinformatics (Jayaram, N., Usvyat, D. and Martin, A.C.R. "Evaluating tools for transcription factor binding site prediction").

2.1 Introduction

2.1.1 Pattern Matching Tools

Pattern matching tools are a key component of TFBS prediction. Pattern matching tools fall into two classes: those that predict individual TFBSs and those that predict clusters of TFBSs.

Pattern matching tools predict putative individual TFBSs or clusters of TFBSs by utilising prior knowledge of the experimentally determined TFBSs describing a transcription factor of interest. These experimentally-determined TFBSs must be represented either as a consensus sequence, or as a PWM (Sand *et al.*, 2008; Elnitski *et al.*, 2006).

A consensus sequence consists of the most frequent base at each position of the experimentally determined TFBS. This consensus sequence can either be strict (using only the 4 letters A, C, G and T from the International Union of Pure and Applied Chemistry (IUPAC) nucleotide code) or degenerate (using the complete 15 letter IUPAC nucleotide code). The complete IUPAC nucleotide code is shown in Table 2.1.

Nucleotide Code	Base
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
R	Adenine or Guanine
Y	Cytosine or Thymine
S	Guanine or Cytosine
W	Adenine or Thymine
K	Guanine or Thymine
M	Adenine or Cytosine
B	Cytosine or Guanine or Thymine
D	Adenine or Guanine or Thymine
H	Adenine or Cytosine or Thymine
V	Adenine or Cytosine or Guanine
N	Any Base

Table 2.1: The complete IUPAC nucleotide code

Use of the strict consensus sequence fails to take important variable regions into account as certain positions within the consensus sequence may consist of nucleotides with equivalent frequencies, therefore, resulting in a more complex pattern. The use of the strict consensus sequence can therefore exclude a subset of the binding site repertoire. Degenerate consensus sequences take into account the occurrence of alternative nucleotides at a

particular position in the TFBS. The use of degenerate consensus sequences characterises the diversity of the TFBS repertoire, and alleviates many of the problems associated with the use of the strict consensus sequence. However, degenerate forms of the consensus sequence fail to take into account the relative frequencies of the alternative nucleotides (Elnitski *et al.*, 2006; Nguyen and Androulakis, 2009; Turatsinze *et al.*, 2008).

The use of PWMs has proven to be very successful in various problems in DNA and protein sequence analysis, and is currently the *de facto* model for TFBS prediction. The PWM model is a matrix of scores which correspond to the frequencies of the four nucleotides at each position in the TFBS motif. In contrast, to the consensus model, the PWM model therefore takes into account the preference for each of the four nucleotides. PWMs can then be visualised as a sequence logo. The fundamental assumption of the PWM model is that the bases at the different positions of the TFBS motif are statistically independent (Nguyen and Androulakis, 2009; Elnitski *et al.*, 2006).

These different ways of representing experimentally determined TFBSs are summarised in Figure 2.1.

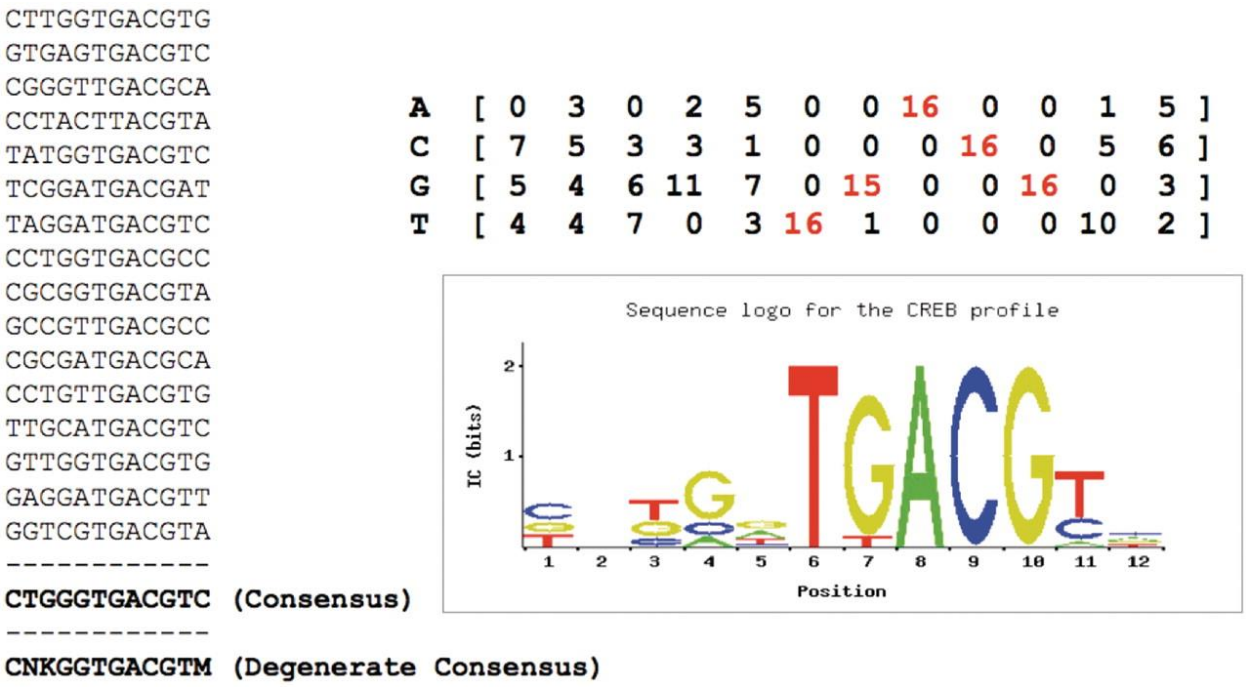


Figure 2.1: Different ways of representing a set of experimentally determined TFBSs (left) including consensus sequences (bottom left),PWMs (top right) and sequence logos(bottom right) (Reproduced from (Zambelli *et al.*, 2012)).

The pattern matching tools use PWMs to predict TFBSs by scanning a DNA sequence of interest with the PWM for a transcription factor of interest. Only one PWM at a time can be used to scan the DNA sequence by the pattern matching tools currently available. The pattern matching tools that predict individual TFBSs scan the DNA sequence in segments which are the same length as the PWM. A raw score or a p-value is calculated to quantify the extent of similarity of the sequence segment to the PWM. The sequence segments which have scores or p-values that exceed a predefined threshold are reported as putative TFBSs (Hannenhalli, 2008; Turatsinze *et al.*, 2008).

The pattern matching tools that predict clusters of TFBSs predict the TFBSs by first scanning the DNA sequence in segments that are the same length as

the PWM. A cluster is defined as a region of DNA that has a high density of predicted TFBSs for a particular transcription factor. The density of a region of DNA in predicted TFBSs is quantified by a raw score or p-value. The regions of DNA that have scores or p-values that exceed a particular threshold are reported as putative clusters of TFBSs (Turatsinze *et al.*, 2008).

The problem with the naïve use of pattern matching tools is that the TFBSs are inherently short and degenerate which can result in a high error rate. There is therefore a need to evaluate the performance of these pattern matching tools in order to improve TFBS prediction (Hannenhalli, 2008; Turatsinze *et al.*, 2008).

2.1.2 Evaluating the Performance of Pattern Matching Tools

In order to conduct an evaluation of pattern matching tools, careful choices need to be made regarding the positive and negative controls and the source of the PWMs used in the evaluation. These are discussed further in the following two sections.

2.1.3 Choice of Positive and Negative Control Sets

Evaluating the performance of pattern matching tools requires positive and negative control sets.

The aim of the positive control is to enable the assessment of performance through calculation of standard performance measures such as sensitivity and positive predictive value. The positive control takes the form of a set of

experimentally characterised TFBSs together with their corresponding gene sequences (Sand *et al.*, 2008).

Experimentally validated precise TFBSs are available either from the commercial TRANSFAC resource, or from the open access resources PAZAR (Portales-Casamar *et al.*, 2009) and ORegAnno (Griffith *et al.*, 2008). Availability and application of the data from TRANSFAC is restricted by a commercial license. Hence, TRANSFAC was rejected for the analyses in this thesis. The data in PAZAR are a superset of ORegAnno, making PAZAR the most comprehensive open access resource. Hence PAZAR has been chosen as the source of the experimentally characterised TFBSs used in the work described in this thesis. PAZAR contains TFBSs for a total of 73 human transcription factors that correspond to a total of 865 unique human genes.

The purpose of the negative control is to assess the false positive rate of a pattern matching tool. The negative control can take the form of artificially generated sequences or randomized gene sequences corresponding to the genome of interest. The problem with using artificially generated sequences is that such sequences are generated using a theoretical background model (e.g. Bernoulli or Markov) which may not reflect the complexity of the genome of interest. Any results obtained can be overly optimistic i.e. result in no TFBSs being predicted by any pattern matching tools and therefore failing to discriminate between pattern matching tools in terms of their false positive rate. This can especially be a problem where vertebrate genomes are concerned given their high level of complexity (Sand *et al.*, 2008). The use of randomized gene sequences relevant to the genome of interest will be more

stringent in terms of the results obtained (Sand *et al.*, 2008). Hence, randomized human gene sequences were chosen to act as the negative control for the work described in this thesis.

2.1.4 Choice of PWM Resource

There are two main resources for obtaining PWMs. These are the commercial resource TRANSFAC, and the open access resource JASPAR. The TRANSFAC resource was established in 1988 and has since been regularly updated. The JASPAR resource was established in 2004 (Sandelin *et al.*, 2004) and has had five further updates. The updates for JASPAR were in 2006 (Vlieghe *et al.*, 2006), 2008 (Bryne *et al.*, 2008), 2010 (Portales-Casamar *et al.*, 2010), 2014 (Mathelier *et al.*, 2013) and 2016 (Mathelier *et al.*, 2015a).

Recently three new open access resources have been established: HOCOMOCO (Kulakovskiy *et al.*, 2013b), HOMER (Heinz *et al.*, 2010), (http://homer.salk.edu/homer/motif/HomerMotifDB/homerResults.html) and CIS-BP (Weirauch *et al.*, 2014). The CIS-BP resource is somewhat different from the other PWM resources; rather than primarily focussing on including newly derived PWMs, the objective is to collate all pre-existing PWMs from existing open-source resources and individual publications and contains some redundancy (i.e. multiple PWMs for a particular transcription factor) (Weirauch *et al.*, 2014).

For the work described in this chapter, the JASPAR resource was chosen to be the source of the PWMs as it is a well-respected freely-available resource that has been available for a long time and is widely used.

2.1.5 Selection of Pattern Matching Tools

A number of pattern matching tools that predict individual TFBSs and clusters of TFBSs have been developed. These pattern matching tools are available in two forms: online and locally-installable. The online forms of the pattern matching tools are only capable of predicting TFBSs for a rather limited number of DNA sequences. Therefore, in order to perform any bulk analysis (e.g. predicting TFBSs within regions identified by ChIP-Seq), the locally-installable forms of the pattern matching tools need to be used. The pattern matching tools that predict individual TFBSs which have a locally-installable version are the open source tools FIMO (Grant *et al.*, 2011), Patser (Turatsinze *et al.*, 2008), Clover (Frith *et al.*, 2004a), PoSSuMsearch (Beckstette *et al.*, 2006) and matrix-scan (Turatsinze *et al.*, 2008), and the commercial tools Match (Kel *et al.*, 2003) and Patch (Matys *et al.*, 2006). The pattern matching tools that predict clusters of TFBSs, and have a locally-installable version are the open source tools MCAST (Bailey and Noble, 2003), BayCis (Lin *et al.*, 2008b), Cister (Frith *et al.*, 2001), Cluster-Buster (Frith *et al.*, 2003), Comet (Frith *et al.*, 2002) and the commercial tool Matrixcatch (Matys *et al.*, 2006).

All of the above pattern matching tools require a set of DNA sequences in FASTA format and PWMs as input and produce a list of putative TFBS clusters, or individual TFBSs, as output. The putative TFBS clusters or individual TFBSs are produced as plain text. Each pattern matching tool requires the PWM to be in a particular file format which differs between different pattern matching tools.

In this evaluation, the decision was taken to use only the pattern matching tools listed above which are open source. Therefore, the pattern matching tools that predict clusters of TFBSs that were chosen for this evaluation were MCAST (Bailey and Noble, 2003), BayCis (Lin *et al.*, 2008b), Cister (Frith *et al.*, 2001), Cluster-Buster (Frith *et al.*, 2003) and Comet (Frith *et al.*, 2002). The pattern matching tools that predict individual TFBSs chosen were FIMO (Grant *et al.*, 2011), Patser (Turatsinze *et al.*, 2008), Clover (Frith *et al.*, 2004a), PoSSuMsearch (Beckstette *et al.*, 2006) and matrix-scan (Turatsinze *et al.*, 2008). Table 2.2 summarises the PWM formats (See section 2.1.6) for each of the pattern matching tools considered in this evaluation and provides the URLs for downloading the pattern matching tools.

TOOL	PWM FORMAT	URL
MCAST	MEME	http://meme-suite.org/
BayCis	tab	http://www.sailing.cs.cmu.edu
Cister	Cluster-Buster	http://zlab.bu.edu
Cluster-Buster	Cluster-Buster	http://zlab.bu.edu
Comet	Cluster-Buster	http://zlab.bu.edu
FIMO	MEME	http://meme-suite.org/
Patser	tab	http://www.rsat.eu/
Clover	Cluster-Buster	http://zlab.bu.edu
PoSSuMsearch	PoSSuM-PSSM	http://bibiserv.techfak.uni-bielefeld.de/
matrix-scan	MEME/Cluster-Buster/TRANSFAC/JASPAR	http://www.rsat.eu/

Table 2.2: Summary of the required PWM formats for each of the pattern matching tools chosen for evaluation and URLs for downloading the tools.

2.1.6 PWM file formats

2.1.6.1 MEME

The pattern matching tools FIMO and MCAST require the PWMs to be in MEME file format. The MEME format illustrated in Figure 2.2 is a plain text format which contains the following sections:

1. the MEME version line which details the oldest version of MEME-SUITE that the file can be read by
2. an alphabet line detailing whether the PWM is for DNA or protein
3. a line giving information on the background frequencies in the source sequence
4. a line giving the name of the motif
5. a line giving information on the alphabet length (this is 4 if the PWM is for a DNA sequence and 20 if it is for a protein sequence) and the width of the PWM
6. A set of records containing the PWM itself. Each row represents the four bases and adds up to one.

```
MEME version 4

ALPHABET= ACGT


Background letter frequencies
A 0.303 C 0.183 G 0.209 T 0.306

MOTIF crp
letter-probability matrix: alength= 4 w= 19
 0.000000  0.176471  0.000000  0.823529
 0.000000  0.058824  0.647059  0.294118
 0.000000  0.058824  0.000000  0.941176
 0.176471  0.000000  0.764706  0.058824
 0.823529  0.058824  0.000000  0.117647
 0.294118  0.176471  0.176471  0.352941
 0.294118  0.352941  0.235294  0.117647
 0.117647  0.235294  0.352941  0.294118
 0.529412  0.000000  0.176471  0.294118
 0.058824  0.235294  0.588235  0.117647
 0.176471  0.235294  0.294118  0.294118
 0.000000  0.058824  0.117647  0.823529
 0.058824  0.882353  0.000000  0.058824
 0.764706  0.000000  0.176471  0.058824
 0.058824  0.882353  0.000000  0.058824
 0.823529  0.058824  0.058824  0.058824
 0.176471  0.411765  0.058824  0.352941
 0.411765  0.000000  0.000000  0.588235
 0.352941  0.058824  0.000000  0.588235
```

Figure 2.2: An example of the MEME format

2.1.6.2 Cluster-Buster

The pattern matching tools Cister, Cluster-Buster, Comet and Clover require the PWM to be in Cluster-Buster format. The Cluster-Buster format shown in Figure 2.3 is a FASTA –like file format for representing PWMs, consisting of a FASTA header line followed by the PWM itself. Each row represents the 4 bases (in the order ACGT) and adds up to one.

```
> crp
 0.000000  0.176471  0.000000  0.823529
 0.000000  0.058824  0.647059  0.294118
 0.000000  0.058824  0.000000  0.941176
 0.176471  0.000000  0.764706  0.058824
 0.823529  0.058824  0.000000  0.117647
 0.294118  0.176471  0.176471  0.352941
 0.294118  0.352941  0.235294  0.117647
 0.117647  0.235294  0.352941  0.294118
 0.529412  0.000000  0.176471  0.294118
 0.058824  0.235294  0.588235  0.117647
 0.176471  0.235294  0.294118  0.294118
 0.000000  0.058824  0.117647  0.823529
 0.058824  0.882353  0.000000  0.058824
 0.764706  0.000000  0.176471  0.058824
 0.058824  0.882353  0.000000  0.058824
 0.823529  0.058824  0.058824  0.058824
 0.176471  0.411765  0.058824  0.352941
 0.411765  0.000000  0.000000  0.588235
 0.352941  0.058824  0.000000  0.588235
```

Figure 2.3: An example of the Cluster-Buster format

2.1.6.3 TRANSFAC

The pattern matching tools Patch, Match and Matrixcatch require the PWMs to be in TRANSFAC format, a plain text format for representing PWMs as shown in Figure 2.4. It contains the following sections:

1. an AC line containing a unique accession code
2. an ID line containing a unique identifier
3. a header row beginning with ‘P0’ containing the order of the bases
4. the PWM itself with each record beginning with a 2-digit position number

Blank rows begin with ‘XX’.

```
AC U00001
XX
ID V$CRP
XX
P0      A         C            G      T
01 0.000000  0.176471  0.000000  0.823529
02 0.000000  0.058824  0.647059  0.294118
03 0.000000  0.058824  0.000000  0.941176
04 0.176471  0.000000  0.764706  0.058824
05 0.823529  0.058824  0.000000  0.117647
06 0.294118  0.176471  0.176471  0.352941
07 0.294118  0.352941  0.235294  0.117647
08 0.117647  0.235294  0.352941  0.294118
09 0.529412  0.000000  0.176471  0.294118
10 0.058824  0.235294  0.588235  0.117647
11 0.176471  0.235294  0.294118  0.294118
12 0.000000  0.058824  0.117647  0.823529
13 0.058824  0.882353  0.000000  0.058824
14 0.764706  0.000000  0.176471  0.058824
15 0.058824  0.882353  0.000000  0.058824
16 0.823529  0.058824  0.058824  0.058824
17 0.176471  0.411765  0.058824  0.352941
18 0.411765  0.000000  0.000000  0.588235
19 0.352941  0.058824  0.000000  0.588235

XX
//
```

Figure 2.4: An example of the TRANSFAC format

2.1.6.4 PoSSuM-PSSM

The pattern matching tool PoSSuMsearch requires the PWMs to be in PoSSuM-PSSM format, a plain text format shown in Figure 2.5. It contains the following sections stored between two lines- BEGIN and END:

1. an ID line containing the identifier for the PWM
2. an AC line containing the accession for the PWM
3. a DE line describing the PWM
4. an AP line detailing whether the PWM is for DNA or protein
5. an LE line specifying the number of rows of the PWM followed by the PWM itself

The order of the bases is introduced by a '#' and the PWM matrix lines start with 'MA'.

```
BEGIN
ID V$CRP
AC U00001
DE CRP
AP DNA
LE 19
#      A        T        C        G
MA 0.000000  0.176471  0.000000  0.823529
MA 0.000000  0.058824  0.647059  0.294118
MA 0.000000  0.058824  0.000000  0.941176
MA 0.176471  0.000000  0.764706  0.058824
MA 0.823529  0.058824  0.000000  0.117647
MA 0.294118  0.176471  0.176471  0.352941
MA 0.294118  0.352941  0.235294  0.117647
MA 0.117647  0.235294  0.352941  0.294118
MA 0.529412  0.000000  0.176471  0.294118
MA 0.058824  0.235294  0.588235  0.117647
MA 0.176471  0.235294  0.294118  0.294118
MA 0.000000  0.058824  0.117647  0.823529
MA 0.058824  0.882353  0.000000  0.058824
MA 0.764706  0.000000  0.176471  0.058824
MA 0.058824  0.882353  0.000000  0.058824
MA 0.823529  0.058824  0.058824  0.058824
MA 0.176471  0.411765  0.058824  0.352941
MA 0.411765  0.000000  0.000000  0.588235
MA 0.352941  0.058824  0.000000  0.588235

END
```

Figure 2.5: An example of the PoSSuM-PSSM format

2.1.6.5 tab

The pattern matching tools Patser and BayCis require the PWMs to be in tab format as shown in Figure 2.6. This represents the PWM as a tab delimited file with a header line introduced by a semi-colon followed by the PWM itself. Each row of the PWM is preceded by A, T, C and G, and separated from the values by a pipe symbol. Each PWM record is ended with two slashes. The tab format has PWM positions across a line and rows representing the bases. This is in contrast to all the preceding matrix formats that have the

four bases going across a line with the rows representing the PWM positions.

```
; MET4 matrix, from Gonze et al. (2005). Bioinformatics 21, 3490-
500.
A |   7   9   0   0  16   0   1   0   0  11   6   9   6   1   8
C |   5   1   4  16   0  15   0   0   0   3   5   5   0   2   0
G |   4   4   1   0   0   0  15   0  16   0   3   0   0   2   0
T |   0   2  11   0   0   1   0  16   0   2   2   2  10  11   8
//
```

Figure 2.6: An example of the tab format

The pattern matching tool matrix-scan is much more flexible regarding the format of the PWMs and accepts PWMs in MEME, Cluster-Buster, tab and TRANSFAC formats as well as the native JASPAR format (which is discussed below).

2.1.6.6 JASPAR

The JASPAR format is a plain text file format shown in Figure 2.7 for representing PWMs. Similarly, to the tab format each row of the PWM is preceded by A, T, C and G and separated from the values by a pipe symbol. The JASPAR format has PWM positions across a line and rows, representing the bases.

```
A|  0  3 79 40 66 48 65 11 65  0
C| 94 75  4  3  1  2  5  2  3  3
G|  1  0  3  4  1  0  5  3 28 88
T|  2 19 11 50 29 47 22 81  1  6
```

Figure 2.7: An example of the JASPAR format

2.1.7 Aim of Chapter

The aim of this chapter is to conduct an independent assessment of a set of pattern matching tools which can be installed locally, and therefore, be used for bulk analysis. This assessment will evaluate the performance of pattern matching tools that predict individual TFBSs and pattern matching tools that predict clusters of TFBSs.

2.2 Methods

All software was locally installed.

There are only 15 human transcription factors which both have PWMs in JASPAR and experimentally characterised binding sites in PAZAR as shown in Figure 2.8. These are BRCA1, E2F1, ELK4, ESR1, ESR2, GATA2, GATA3, IRF1, MAX, NFKB, STAT1, YY1, CTCF, NF-YA and SP1. Hence, the performance of the pattern matching tools could only be assessed for these 15 transcription factors.

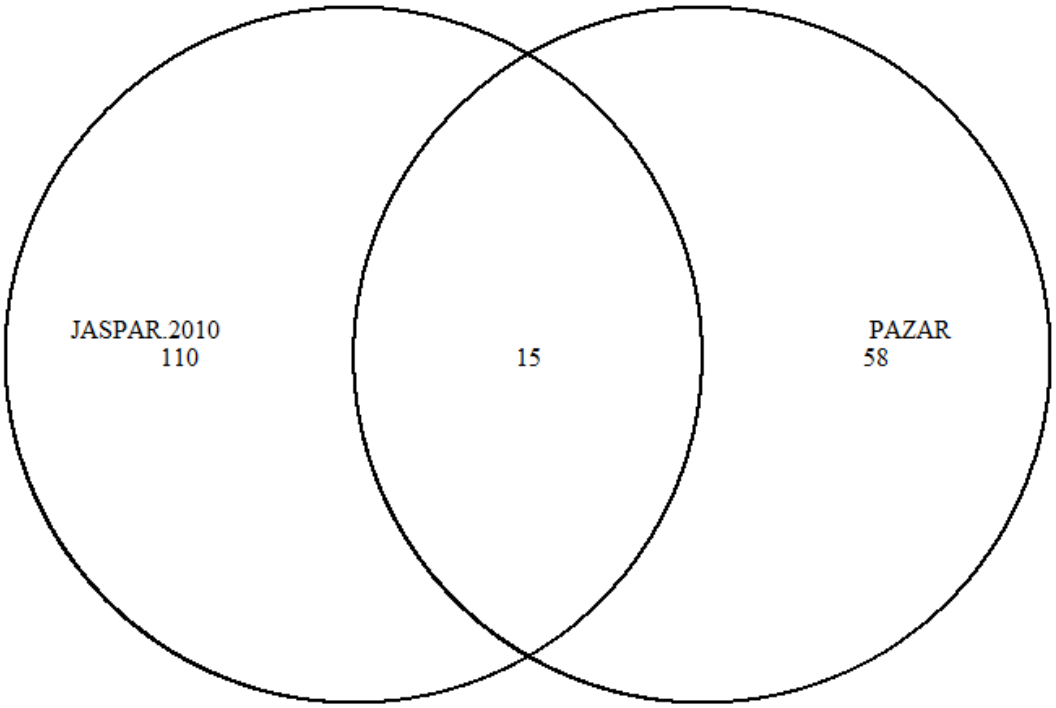


Figure 2.8: Venn diagram showing the overlap between the PWMs in JASPAR.2010 and the experimentally characterised TFBSs in PAZAR

The PWMs for the 15 human transcription factors were obtained from the 2010 release of JASPAR (JASPAR.2010) (Portales-Casamar *et al.*, 2010), as this was the latest available release at the time this work was carried out. These PWMs are derived from SELEX, or individual promoter assays, and were in the JASPAR format.

Of all the tools chosen for evaluation, only matrix-scan can accept the JASPAR format. For the remainder of the pattern matching tools, these PWMs were converted from the JASPAR format into the formats required for the particular pattern matching tool using the convert-matrix program from the RSAT suite (Thomas-Chollier *et al.*, 2011) to generate Cluster-Buster and the tab formats, and the jaspar2meme program, from MEME-SUITE (Bailey *et al.*, 2015) to generate the MEME format. There is no single program capable of converting the JASPAR format to the PoSSuM-PSSM format, so the convert-matrix program was used to convert the PWMs from JASPAR to TRANSFAC format and then the transfac2gen program (which is included with the PoSSuMsearch download) was used to convert the PWMs from TRANSFAC to PoSSuM-PSSM format.

2.2.1 Evaluating Performance

Known precise TFBSs, experimentally characterized from biochemical protein-DNA binding experiments, corresponding to the 15 human transcription factors, were downloaded from PAZAR in GFF format (Portales-Casamar *et al.*, 2009), for a total of 181 human genes. The GFF format shown in Figure 2.9 is a tab separated file format used for storing genomic information. The first line of the file, consists of a comment identifying the file format, and version. This is followed by a set of lines describing the data. Each line contains the following fields:

1. the chromosome name
2. the source name
3. the feature name
4. the start position
5. the end position
6. the score of the feature
7. the strand
8. the attributes (a semicolon separated list of key-value pairs that provide additional information)

A '.' character is used to represent any empty fields.

```
##gff-version 3
1   .   TFBS   1300   1315   .   +   ID=TFBS00001
1   .   TFBS   1050   1060   .   +   ID=TFBS00002
1   .   TFBS   3000   3012   .   +   ID=TFBS00003
1   .   TFBS   5000   5014   .   +   ID=TFBS00004
1   .   TFBS   7000   7009   .   +   ID=TFBS00005
```

Figure 2.9: Example of the GFF format

PAZAR contains some redundancy (multiple instances of the same TFBS annotated for a given gene at the same location), so, any duplicated TFBSs were removed using the UNIX command uniq. Subsets of this dataset were then selected, which contained at least one TFBS for the transcription factor being evaluated in a particular comparison. The GFF file was then converted into BED format using the GFF-to-BED conversion utility in Galaxy (Goecks *et al.*, 2010). The BED format shown in Figure 2.10 is also a tab delimited format used for storing genomic information containing the following compulsory fields:

1. chromosome
2. start position
3. end position

```
chr7    127471196  127471211
chr7    127472363  127472375
chr7    127473530  127473540
chr7    127474697  127474706
chr7    127475864  127475877
chr7    127477031  127477042
chr7    127478198  127478212
chr7    127479365  127479374
chr7    127480532  127480542
```

Figure 2.10: An example of the BED format

TFBSs can occur in the promoter region, and in introns and exons, as well as far upstream of genes (up to 10,000 bp) (Farnham, 2009; Cline and Karchin, 2011). Consequently, the complete gene sequence (i.e. both exons and introns), together with an upstream region of 10,000 bp of each of the 181 genes was obtained in FASTA format. In addition, the genomic coordinates of these sequences were obtained as a text file. The DNA sequences and genomic coordinates were obtained from Biomart (Smedley *et al.*, 2009) using the biomaRt package in Bioconductor (Durinck *et al.*, 2005; Durinck *et al.*, 2009; Gentleman *et al.*, 2004). The genomic coordinates were converted to BED format using Pybedtools (Dale *et al.*, 2011).

Prediction of TFBSs was carried out using the selected pattern matching tools together with the PWMs using the DNA sequences obtained from Biomart. The pattern matching tools were run at their default cutoff

thresholds, as this is normal practice in order to minimise the false positives and false negatives while maximising the true positives (Sand *et al.*, 2008; Turatsinze *et al.*, 2008). All of the resulting text files containing the predicted TFBSs were converted to BED format using Pybedtools (Dale *et al.*, 2011).

The coordinates of the predicted TFBSs from all of the selected pattern matching tools are relative to their larger genomic fragments (i.e. they are relative coordinates). The coordinates of the experimentally characterised TFBSs obtained from PAZAR are genomic coordinates (i.e. describing their actual location in the genome). In order to compare the predicted TFBSs and the known TFBSs, the coordinates of the predicted TFBSs were converted from relative to genomic coordinates, using the convert-feature program from RSAT (Thomas-Chollier *et al.*, 2011) with output in BED format. The genomic coordinates of DNA sequences obtained previously, were provided as the source of genomic coordinates to the convert-feature program. The convert-feature program requires all input files in BED format, so both the coordinates of the predicted TFBSs and the coordinates of the DNA sequences obtained from Biomart were first converted to BED format.

The predicted TFBSs were compared with the experimentally characterised TFBSs, using the intersectBed program from the BEDTools suite (Quinlan and Hall, 2010), which requires all input files to be in BED format. Hence the reason for converting the coordinates of the experimentally characterised TFBSs obtained from PAZAR to BED format.

True positives were defined as predicted binding sites having a minimum overlap of 70% of base pairs with known binding sites from PAZAR. Similarly, false positives were defined as predicted binding sites not having an overlap of at least 70% of base pairs with a known binding site, and false negatives were defined as known binding sites that were not identified. The overlap of 70% in the context of the evaluation of performance of pattern matching tools is a practice that has been recommended by Sand *et al.* (2008). Obtaining a true estimate of the total number of negative sites (and hence the number of true negatives) is hard due to the ambiguities that exist in defining negative sites that are neither experimentally characterised nor computationally predicted. Therefore the normal practice of avoiding performance measures that require true negative counts was adopted (Sand *et al.*, 2008). This comparison of the experimentally characterised TFBSs with the predicted TFBSs is shown in Figure 2.11.

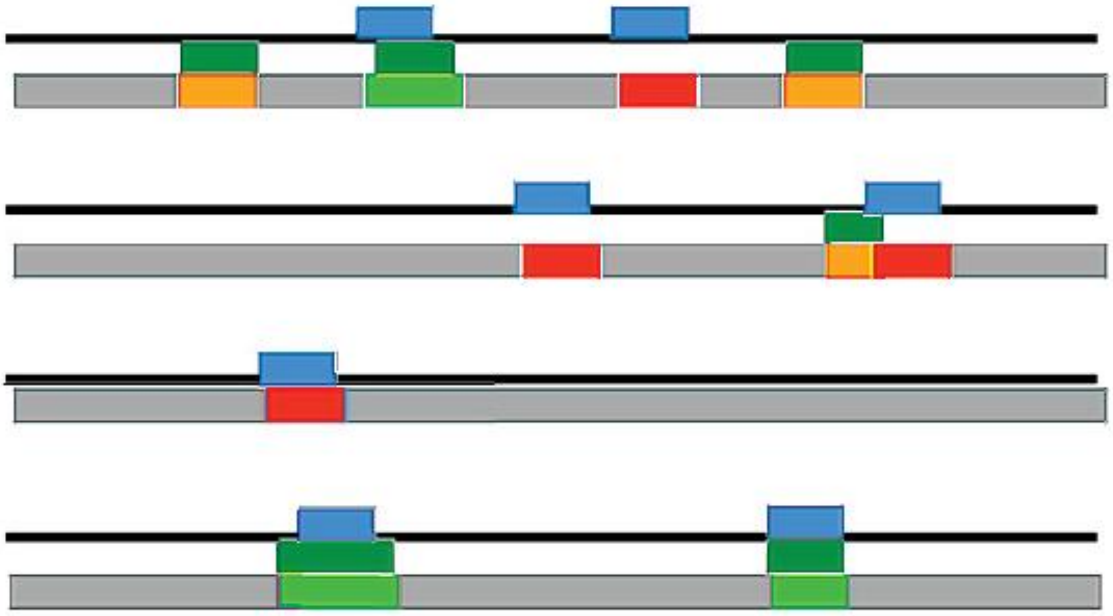


Figure 2.11: A schematic illustration of the comparison between known and predicted TFBSs which are represented by blocks. The dark green blocks represent the known TFBSs, the blue blocks represent the predicted TFBSs, the light green blocks represent true positives, the red blocks represent false positives and the orange blocks represent false negatives (Reproduced from (Sand *et al.*, 2008)).

For pattern matching tools that predict clusters of TFBSs, all predicted component TFBSs within a region were required to overlap with experimentally characterised sites, by a minimum of 70% of base pairs, for a prediction to be regarded as a true positive.

Performance was assessed by calculating sensitivity, positive predictive value and geometric accuracy. These were averaged across the transcription factors and genes analysed.

The sensitivity (Sn) describes the fraction of the experimentally characterised TFBSs that are covered by the predicted TFBSs which is calculated as

$$Sn = \frac{TP}{(TP + FN)} \quad (2.1)$$

Where TP is the number of true positives and FN is the number of false negatives.

The positive predictive value (PPV) describes the fraction of the predicted TFBSs that are also found in the set of experimentally characterised TFBSs. It is calculated as

$$PPV = \frac{TP}{(TP + FP)} \quad (2.2)$$

Where TP is the number of true positives and FP is the number of false positives.

The geometric accuracy (ACCg) describes the trade-off between the sensitivity and positive predictive value and is calculated as

$$ACCg = \sqrt{Sn \times PPV} \quad (2.3)$$

In order to create the randomized gene sequences for the negative control, all the DNA sequences were scrambled using the shuffleseq program from the EMBOSS suite (Rice *et al.*, 2000).

In the case of the TFBSs predicted using scrambled sequences, there are no actual positives and therefore no true positives or false negatives. Any predictions are therefore classified as false positives. Performance was thus assessed for the scrambled sequences by calculating false positive rate which is described here as the 'false positive rate on scrambled sequences' (FPRs). The above steps are summarised in Figure 2.12.

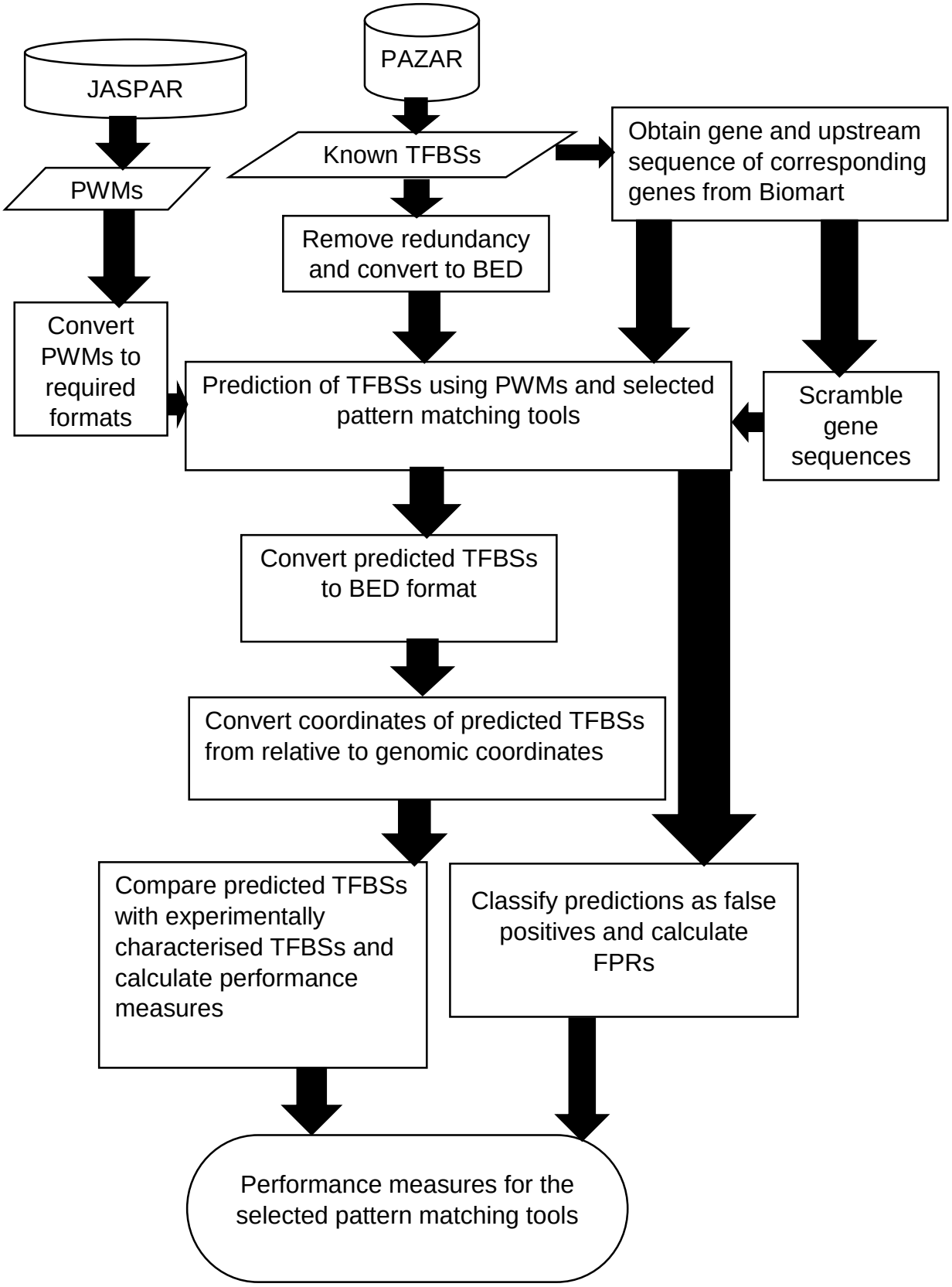


Figure 2.12: Flowchart summarising methods to evaluate performance of pattern matching tools. See Text.

The FPRs is calculated as

$$FPRs = \frac{Np}{AN} \quad (2.4)$$

Where Np is the number of predicted sites and AN is the number of actual negatives.

As discussed in section 2.2.1, obtaining the number of actual negatives (AN) (and specifically the number of true negatives) is a hard problem. AN is normally calculated as

$$AN = FP + TN \quad (2.5)$$

Where FP is the number of false positives and TN is the number of true negatives

In calculating FPRs, the AN was defined as

$$AN = \frac{L}{l_t} \quad (2.6)$$

Where L is the length of the DNA sequence and l_t is the length of the PWM in question.

The FPRs was averaged across the transcription factors and genes analysed.

2.3 Results and Discussion

Table 2.3 shows that FIMO (Grant *et al.*, 2011) and MCAST (Bailey and Noble, 2003) are the best performing pattern matching tools that predict individual TFBSs and clusters of TFBSs respectively. In general, the pattern matching tools predicting individual TFBSs perform better than those predicting clusters of TFBSs. Because of the more stringent requirements for a true positive in predicting clusters of TFBSs (i.e. every predicted site within the cluster must have a minimum 70% overlap with a true site), it might be expected that the sensitivity for pattern matching tools that predict clusters of TFBSs would be lowered, while the specificity would be improved. Indeed, the sensitivity of the pattern matching tools that predict clusters of TFBSs is somewhat lower than the pattern matching tools that predict individual TFBSs. Since the true negative count is not available, the specificity cannot be calculated, but surprisingly the FPR_s for the pattern matching tools that predict clusters of TFBSs is larger than that for the pattern matching tools that predict individual TFBSs suggesting that the pattern matching tools that predict clusters of TFBSs have lower specificity.

While the pattern matching tools which tools that predict individual TFBSs outperform those that predict clusters of TFBSs, the choice of the type of pattern matching tool depends on the context in which it is to be used. Consequently, if prior knowledge is available about the DNA sequence being scanned (i.e. a ChIP-Seq region) then using a pattern matching tool that predicts individual TFBSs is probably a sensible strategy. When analysing a stretch of DNA with no prior knowledge about the presence of a gene, it would be better to use a prediction tool that identifies clusters of TFBSs since the chance of a random match is much reduced.

	Sn	PPV	ACCg	FPRs
Individual				
FIMO	0.815	0.735	0.774	0.015
Patser	0.722	0.653	0.687	0.016
PoSSuMsearch	0.708	0.635	0.670	0.020
Clover	0.673	0.584	0.627	0.023
matrix-scan	0.647	0.579	0.612	0.028
Cluster				
MCAST	0.774	0.683	0.727	0.033
BayCis	0.598	0.497	0.545	0.040
Cister	0.635	0.565	0.599	0.040
Cluster-Buster	0.657	0.581	0.617	0.039
Comet	0.682	0.589	0.634	0.038

Table 2.3: Performance of the selected pattern matching tools using PWMs from JASPAR.2010. Average sensitivities (Sn), Positive Predictive Value (PPV) and geometric accuracy (ACCg) are reported together with the false positive rate using scrambled sequences (FPRs). Performance evaluation was performed across the 15 transcription factors that overlap between PAZAR and JASPAR.

.

2.4 Conclusions

As a comprehensive set of experimentally-characterized transcription factor binding sites is not available, having good reliable prediction methods is very important. While the need for these as an adjunct to gene prediction in the human genome has diminished owing to the wide scale experimental characterisation of transcription factor binding via high-throughput ChIP experiments, it is now much more important in order to have a full understanding of the regulation of gene expression, and to be able to consider the potential phenotypic effects of mutations occurring in a TFBS.

However, these high-throughput ChIP experiments do not identify the precise TFBS; therefore, in order to make full use of the experimental maps of transcription factor binding, the precise TFBS must still be identified within a much wider window of bases. This needs to be done by computational prediction.

Evaluating the performance of the pattern matching tools has the potential to improve the computational prediction of TFBSs, and hence, aid the analysis and interpretation of data from large scale sequencing projects.

In this chapter, a set of transcription factor binding site prediction tools that could be downloaded and installed locally have been evaluated, identifying FIMO and MCAST as the best-performing tools for identifying individual TFBSs and clusters of TFBSs respectively.

3 An Independent Assessment of Motif Discovery Tools

Parts of the work presented in this chapter was presented as a poster at the MASAMB 2013 conference held at Imperial College London and has been submitted as a paper to BMC Bioinformatics (Jayaram, N., Usvyat, D. and Martin, A.C.R. "Evaluating tools for transcription factor binding site prediction").

3.1 Introduction

In the previous chapter, the performance of a set of pattern matching tools in TFBS prediction was evaluated. The tool FIMO was found to be the best performing. However, in addition to having a pattern matching tool with as high a performance as possible, the computational prediction of TFBSs requires a set of high quality PWMs.

This chapter discusses the evaluation of a set of motif discovery tools, the identification of the best performing motif discovery tool (rGADEM), the creation of a new set of PWMs using rGADEM, and the re-evaluation of the pattern matching tools that were evaluated in the previous chapter using the

new PWMs, to ensure that PWM choice does not have a major impact on the performance of the pattern matching tools.

3.1.1 De Novo Motif Discovery

PWM models are derived by *de novo* motif discovery from a set of TFBSs that have been experimentally determined to bind a particular transcription factor. This is done by using one of several *de novo* motif discovery programs which identify a common over-represented signature, or motif, and derive a PWM for the transcription factor (Narlikar and Ovcharenko, 2009). Motif discovery tools exist in both online and locally-installable forms. The online forms of the motif discovery tools are only capable of deriving PWMs for a very limited number of DNA sequences. Therefore, in order to perform any bulk analysis, the locally-installable forms of the motif discovery tools need to be used.

A plethora of classical *de novo* motif discovery tools (i.e. deriving PWMs from a set of TFBSs collated from SELEX or individual promoter assays) have been developed. The classical *de novo* motif discovery tools that have a locally-installable version are: AlignAce (Hughes *et al.*, 2000), Consensus (Hertz and Stormo, 1999), GLAM (Frith *et al.*, 2004b), The Improbizer (Ao *et al.*, 2004) , MEME (Bailey and Elkan, 1994), MotifSampler (Thijs *et al.*, 2001) and SesiMCMC (Favorov *et al.*, 2005).

The large volumes of data generated from the high-throughput techniques ChIP-chip and ChIP-Seq have presented challenges to *de novo* motif discovery. For example, a ChIP-Seq experiment can generate over 10,000

sequences in a single run. However, the conventional *de novo* motif discovery programs were developed when only a small number of protein-DNA binding events could be characterised, and as such, are not equipped to handle large volumes of data.

Hence a common practice has been to use these tools on a subset of the sequences (Jothi *et al.*, 2008; Valouev *et al.*, 2008; Hu *et al.*, 2010). However, Hu *et al.* (2010) have suggested that this practice will lead to inaccurate PWMs. Therefore, the tools ChIPMunk (Kulakovskiy *et al.*, 2010), HOMER (Heinz *et al.*, 2010), rGADEM (Mercier *et al.*, 2011) and MEME-ChIP (Ma *et al.*, 2014; Machanick and Bailey, 2011) have recently been developed that are able to handle the large volumes of data generated from these high-throughput technologies. All of these tools have a locally-installable version available.

3.1.2 Impact of High-Throughput Technologies on Motif Discovery

It has been suggested that PWMs derived from data from the high-throughput techniques ChIP-chip and ChIP-Seq methods, will be more accurate than PWMs derived from data from techniques such as SELEX, or compilations of individual promoter assays that detect limited transcription factor binding site numbers. Furthermore, the ChIP-Seq technique has been found to produce PWMs with greater accuracy than ChIP-chip owing to the superior resolution provided by the ChIP-Seq technique (Hu *et al.*, 2010; Portales-Casamar *et al.*, 2010).

3.1.3 Aim of Chapter

The aims of this chapter are firstly to conduct an independent assessment, using the ENCODE ChIP-Seq data, of the locally-installable motif discovery tools that are able to handle large volumes of data; Secondly, to generate a set of PWMs using these data with the best-performing motif discovery tool; Finally, to check that the selection of the best pattern matching tool is not unduly influenced by the choice of PWMs.

3.2 Methods

All software was locally installed.

3.2.1 Overlap between Resources

There are currently a total of 90 transcription factors that are represented in the ENCODE ChIP-Seq data. ChIP-Seq datasets for 29 transcription factors have access restrictions. Only the transcription factors that had ChIP-Seq datasets with no access restrictions were selected.

The 61 transcription factors without access restrictions are: AP-2A, AP-2Y, ATF3, BHLHE40, BRCA1, BRF2, CHD2, C-FOS, C-JUN, C-MYC, CEBPB, CTCF, E2F1, E2F4, E2F6, EBF1, ELK4, ERRA, GATA1, GATA2, GATA3, GRP20, GTF2B, HA-E2F1, HNF4A, HSF1, IRF1, IRF3, JUND, KAP1, MAFF, MAFK, MAX, NF-E2, NF-YA, NF-YB, NFKB, NRF1, POL2, PRDM1, RFX5, RPC155, SETDB1, SPT20, SREBP1, SREBP2, STAT1, STAT2, STAT3, TAL1, TBP, TCF7L2, TFIIIC-110, TR4, USF2, YY1, ZNF143, ZNF217,

ZNF263, ZNF274 and ZZZ3. However, in order to perform an independent assessment of the different motif discovery tools and to check that the selection of the best pattern matching tool is not unduly influenced by the choice of PWMs, the transcription factors will need to have experimentally characterised TFBSs in PAZAR.

Out of the selected 61 transcription factors, there are only 13 human transcription factors which are represented in the ENCODE ChIP-Seq data and have experimentally characterised TFBSs in PAZAR (Portales-Casamar *et al.*, 2009) as shown in Figure 3.1. These are BRCA1, E2F1, ELK4, GATA2, GATA3, IRF1, MAX, NFKB, STAT1, YY1, CTCF, NF-YA and TAL1. Hence, unless otherwise specified, the evaluations discussed in this chapter could only be performed for these 13 transcription factors.

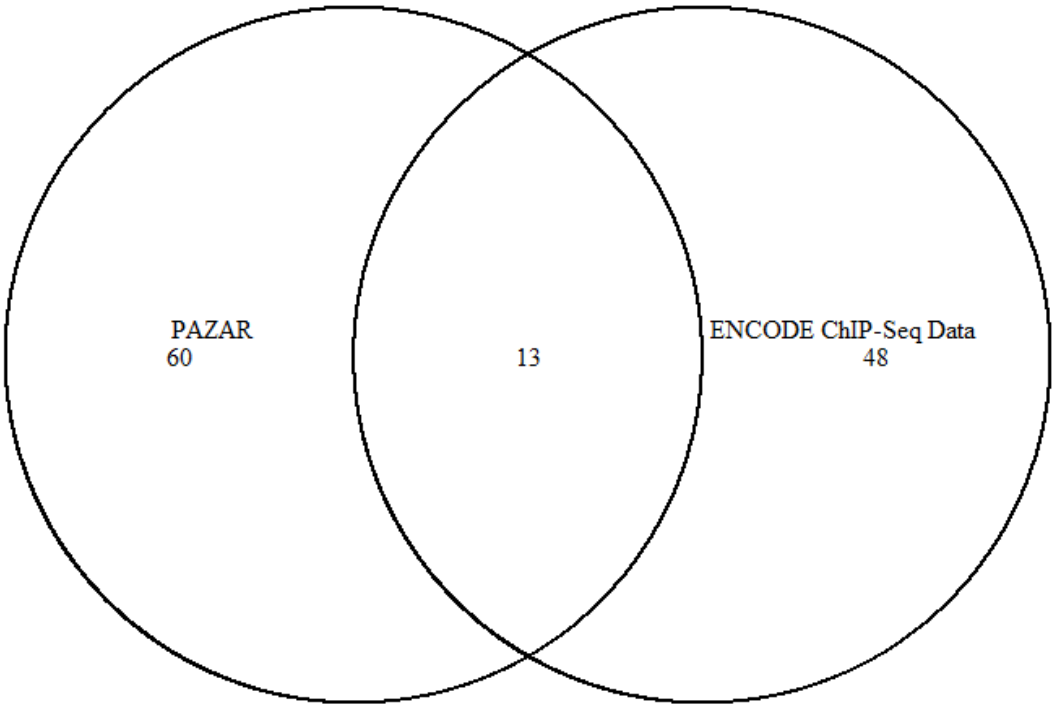


Figure 3.1: Overlap of transcription factor data.

The Venn diagram shows overlaps between experimentally characterised TFBSs in PAZAR, and those transcription factors represented in the ENCODE ChIP-Seq data.

3.2.2 Deriving PWMs

The methods used for deriving *de novo* PWMs are summarized in Figure 3.2.

For each transcription factor represented in the ENCODE project, two sets of ChIP-Seq samples together with a ChIP-Seq control sample are available for

each transcription factor. ChIP-Seq control samples are obtained from a mock experiment without the specific antibody (Bardet *et al.*, 2012).

All of the ChIP-Seq datasets for the selected human transcription factors were downloaded from the ENCODE project (http://hgdownload.cse.ucsc.edu/goldenPath/) in FASTQ format (see section 1.2.2). It is important that the short reads arising from ChIP-Seq are aligned properly to the reference genome, otherwise false positives and false negatives would arise from the reads being mapped to the wrong location. In order to avoid this, the quality of the reads was checked using FastQC (http: //www.bioinformatics. babraham.ac.uk/projects/fastqc/). Any adaptors and low quality reads were then removed using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). This was done for all of the ChIP-Seq datasets for the selected human transcription factors.

The reads were then mapped to the human genome version hg19 using Bowtie (Langmead *et al.*, 2009). This was done for all of the ChIP-Seq datasets for the selected human transcription factors.

Bowtie was chosen because it is the recommended aligner for ChIP-Seq data (Bardet *et al.*, 2012). The –best parameter was used so that the best alignment for a particular read would be reported. This however tends to reduce the speed. The –m parameter was set to 1 to ensure that only reads that aligned to one part of the genome were aligned. This is a practice that is recommended by Bardet *et al.* (2012). The –n parameter controls the allowed number of mismatches between the read and the genome would be allowed and ranges between 0 and 2. Upon investigation of different values

from 0 to 2 for –n, a value of 1 was found to give the highest percentage of reads that aligned to the genome without ambiguity and was used for all experiments. The hg19 version of the human genome, the latest available release at the time of doing this work was downloaded from ftp://ftp.ccb.jhu.edu/pub/data/bowtie_indexes/hg19.ebwt.zip. The resulting Sequence Alignment/Map format (SAM) files (see section 1.2.2.2) were converted to binary format (BAM) files, and indexed using SAMtools (Li *et al.*, 2009b). The BAM format is the binary version of the SAM format. This step reduces the file size, and allows rapid access which is essential given the large size of the data (several gigabytes). These BAM files were then converted to BED format (see section 2.2.1) using the bamtobed program in the BEDTools suite (Quinlan and Hall, 2010). These steps were done for all of the ChIP-Seq datasets for the selected human transcription factors.

After the reads were aligned to the reference genome, peak calling was performed by identifying statistically significant binding regions that are enriched in the ChIP-Seq sample compared with the control sample (Park, 2009). The use of the control sample in the peak calling step helps to control biases and artefacts that occur in the experimental protocol (Park, 2009; Bardet *et al.*, 2012) as recommended by Bardet *et al.* (2012). Peaks were called using MACS (Zhang *et al.*, 2008) for both ChIP-Seq samples for each transcription factor in the set of selected transcription factors. MACS was chosen as it is the recommended peak caller for calling peaks that are to be used in *de novo* motif discovery (Wilbanks and Facciotti, 2010). Default parameters were used as this is the practice recommended by Wilbanks and Facciotti (2010) and Bardet *et al.* (2012). MACS requires the input to be in

BED format hence the reason for converting the BAM files to BED format. Common peaks between ChIP-Seq samples for a particular transcription factor were selected for further analysis, using the Bioconductor package ChIPpeakAnno (Zhu *et al.*, 2010; Gentleman *et al.*, 2004). This is a practice that has been recommended by Bailey *et al.* (2013). A set of peak regions, centred on the summits of the peaks (±100 bp), were obtained, in order to prevent bias towards longer peak regions (Bardet *et al.*, 2012). These peak regions were then converted to FASTA format using the Bioconductor package ChIPpeakAnno (Zhu *et al.*, 2010; Gentleman *et al.*, 2004). This is because the motif discovery tools require DNA sequences in FASTA format as input.

For evaluation purposes, *de novo* motif discovery was carried out on the peak regions derived from the ENCODE ChIP-Seq data using MEME-ChIP (Ma *et al.*, 2014; Machanick and Bailey, 2011), HOMER (Heinz *et al.*, 2010), ChIPMunk (Kulakovskiy *et al.*, 2010) and rGADEM (Mercier *et al.*, 2011) for the 13 transcription factors that overlap between PAZAR and the ENCODE ChIP-Seq data. Since these programs are able to deal with large datasets, all peak regions were used.

3.2.3 Finding optimum parameters for the motif discovery tools

It is key that the PWM generated for a particular transcription factor matches the experimentally validated binding pattern in the literature. The tools have

parameters that can be adjusted for motif discovery, and all possible combinations of these were explored, using a step size of 10%, in order to generate PWMs that resembled the experimentally validated binding pattern in the literature. It was found that for MEME-ChIP, HOMER and ChIPMunk the default values, for the parameters produced PWMs that resembled the experimentally validated binding patterns. Any change from the default values produced PWMs that were drastically different from these experimentally validated binding patterns. In the case of rGADEM however, the e-value parameter had to be set to a value of 0.5, with the remainder of the parameters set at their default values, to generate PWMs that resembled experimentally validated binding patterns. Again, deviation from these values including use of the default e-value of 0.0, resulted in PWMs that were drastically different from the experimentally validated binding patterns. During the exploration of parameters, it was found that the first PWM generated, always best resembled the experimentally validated binding pattern, and consequently the motif discovery tools were set to generate just one PWM for a particular transcription factor.

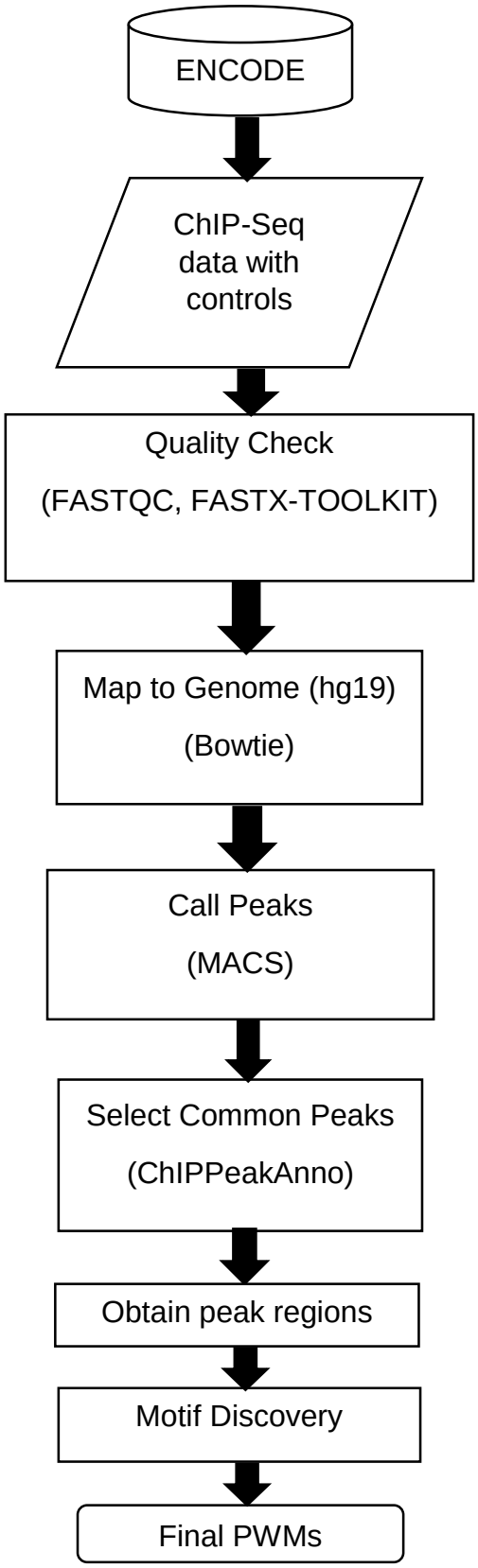


Figure 3.2: Flowchart summarising methods used to derive PWMs from the ENCODE ChIP-Seq data. See text.

3.2.4 Evaluation of Motif Discovery Methods

Logically, it makes sense, to evaluate motif discovery methods first, and then, to evaluate the tools available for matching the derived PWMs to DNA sequences. However, the evaluation of the performance of motif discovery methods requires a pattern matching tool to test the performance of the resulting PWMs. Therefore, a pattern matching tool must be selected for this purpose. In the previous chapter, the performance of a number of pattern matching tools in TFBS prediction was evaluated using PWMs from the 2010 release of JASPAR. The best performing tool (FIMO) was then used in this chapter for evaluating the motif discovery methods.

The FIMO pattern matching tool requires the input PWMs in MEME format. The motif discovery tool MEME-ChIP generates PWMs in this format, while the rGADEM, ChIPMunk and HOMER motif discovery tools all produce PWMs in tab format. There is no single program capable of converting PWMs from tab to MEME format. Therefore, the PWMs that were derived using rGADEM, ChIPMunk and HOMER were first converted from tab to JASPAR format using the convert-matrix program from the RSAT suite (Thomas-Chollier *et al.*, 2011). These were then converted from JASPAR to MEME format using the jaspar2meme program , from MEME-SUITE (Bailey *et al.*, 2015).

Performance was evaluated on the PWMs that resembled well established motifs by using the FIMO motif scanning tool using the protocol outlined in section 2.2.1 with the exception that since the evaluation was performed

over the 13 transcription factors that overlapped between PAZAR and ENCODE ChIP-Seq the number of genes used in the evaluation was 167.

The PWMs obtained using the different methods were compared with each other in terms of their similarity. This was done by the calculation of normalised Euclidean distances between equivalent PWMs using the TFBSTools package in Bioconductor (http://www.bioconductor.org/packages/release/bioc/html/TFBSTools.html). The normalised Euclidean distance was chosen as it has been found to be the most effective method for comparing PWM similarity (Gupta *et al.*, 2007). Reverse complement matrices were also checked, and the minimum distances recorded. Results for each matrix set comparison were averaged across the transcription factors used. The normalised Euclidean distance ranges from 0 to 1 where 0 denotes complete identity and 1 denotes complete dissimilarity. The normalised Euclidean distance is calculated as follows:

$$D(a,b) = \frac{1}{2l} \cdot \sum_{i=1}^{l} \sqrt{\sum_{b \in \{A,C,G,T\}} \left(p_{i,b}^{1} \_ p_{i,b}^{2}\right)^2} \qquad (3.1)$$

Where l is the length of the PWM, $p_{i,b}^{1}$ is the value in column i with DNA base b for PWM 1 and $p_{i,b}^{2}$ is the value in column i with DNA base b for PWM 2.

3.3 Results and Discussion

3.3.1 Evaluating the Performance of Motif Discovery Tools

Table 3.1 shows that rGADEM has the best performance and MEME-ChIP the worst on all four performance measures.

Table 3.2 shows that by comparing the PWMs generated in this work using different motif discovery tools, that while all PWMs resemble experimentally validated binding patterns, there are significant differences in the PWMs generated by different motif discovery tools.

Table 3.2 also shows that the largest difference is between the PWMs derived using the best performing method (rGADEM) and PWMs from the worst performing method (MEME-ChIP).

Figure 3.3 shows the similarity between the PWMs generated by the different motif discovery tools in the form of a tree. This was created by using Table 3.2 as vectors for clustering using Ward's minimum variance technique and visualised using the DRAWGRAM utility from the Phylip package (http://evolution.genetics.washington.edu/phylip.html).

TOOL	Sn	PPV	ACCg	FPRs
ChIPMunk	0.886	0.786	0.835	0.009
MEME-ChIP	0.865	0.771	0.817	0.012
rGADEM	0.932	0.840	0.885	0.002
HOMER	0.901	0.794	0.846	0.007

Table 3.1: Performance of the different motif discovery tools using FIMO.

Average sensitivities (Sn), Positive Predictive Value (PPV) and geometric accuracy (ACCg) and false positive rate on scrambled sequences (FPRs) are reported. Note that PWMs were generated only for the 13 transcription factors that overlap between the ENCODE ChIP-Seq data and PAZAR data.

	rGADEM	HOMER	ChIPMunk	MEME-ChIP
rGADEM	0	–	–	–
HOMER	0.161	0	–	–
ChIPMunk	0.264	0.120	0	–
MEME-ChIP	0.372	0.202	0.153	0

Table 3.2: Normalised Euclidean distances between PWMs derived using the different motif discovery tools. Note that comparisons between the matrices generated in this work were performed over the 13 PWMs which were used for performance evaluation (i.e. those that correspond to transcription factors that overlap between ENCODE-ChIP-Seq data and PAZAR).

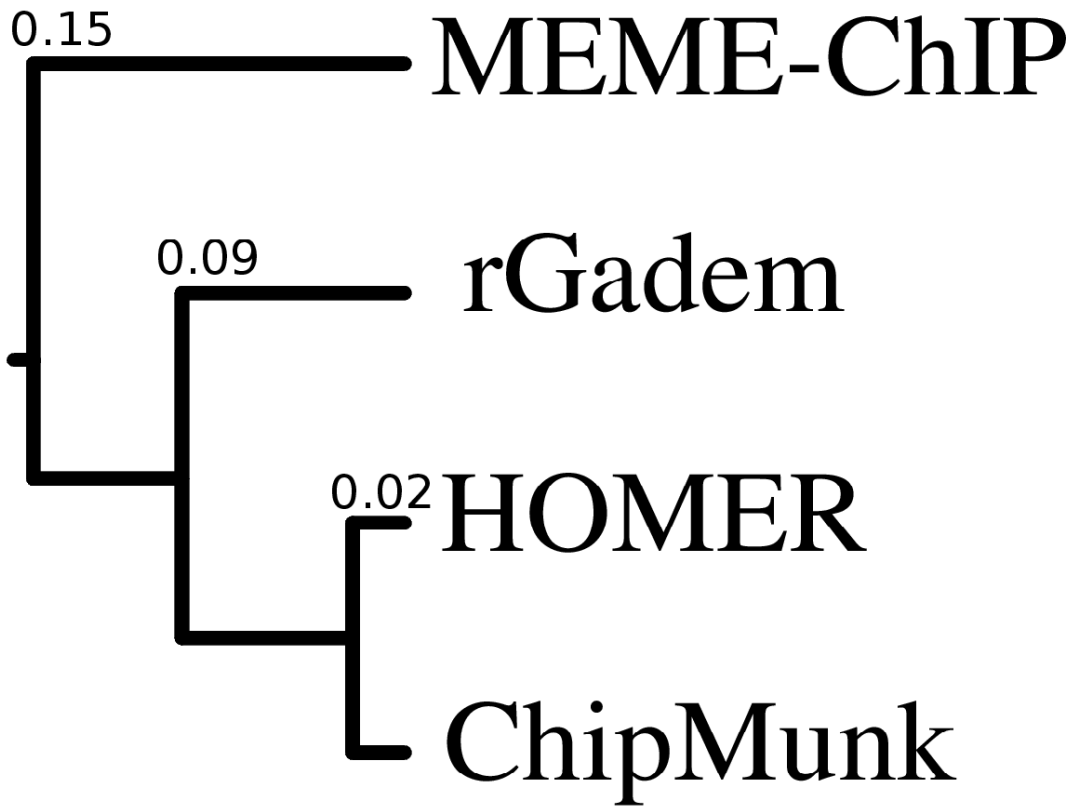


Figure 3.3: Tree showing the similarity between the PWMs generated by the different motif discovery tools

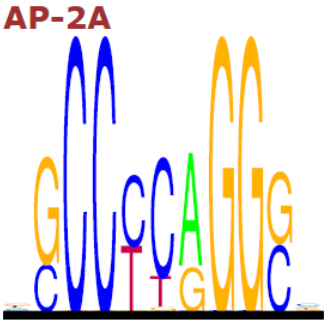
3.3.2 Derivation of a New Set of PWMs

Having shown that the rGADEM motif discovery method clearly out-performs the other methods, motif discovery was then performed on a further set of 48 transcription factors, using rGADEM on the peak regions derived from the ENCODE ChIP-Seq data. It was decided to store the PWMs in MEME format as this is the format required for FIMO which is the best performing pattern matching tool. There is no single program capable of converting PWMs from tab to MEME format. Therefore the resulting PWMs were converted from the tab to JASPAR format using the convert-matrix program from the RSAT suite

(Thomas-Chollier *et al.*, 2011) and then from JASPAR to MEME format using the jaspar2meme program , from MEME-SUITE (Bailey *et al.*, 2015).

3.3.3 The hCRM Resource

The 61 PWMs derived using rGADEM were then made publically available via the web. This new resource is referred to hereafter as the 'human ChIP-Seq rGADEM matrices' (hCRM) and they may be downloaded from http://www.bioinf.org.uk/tfbs/. The hCRM PWMs can be obtained individually or via bulk download as a ZIP file or gzipped tar file. This web site also allows the matrices to be browsed and viewed as 'sequence logos', and downloaded individually. The two ways the website can be displayed are summarised in Figure 3.4 and Figure 3.5.



Background letter frequencies

A 0.25 **C** 0.25 **G** 0.25 **T** 0.25

Motif width: 11

A	C	G	T
0.2481	0.1469	0.3298	0.2752
0.0005	0.3046	0.6944	0.0005
0.0005	0.9985	0.0005	0.0005
0.0005	0.9985	0.0005	0.0005
0.0005	0.5175	0.0005	0.4815
0.0005	0.7719	0.0240	0.2036
0.7249	0.0005	0.2741	0.0005
0.0005	0.0005	0.9985	0.0005
0.0005	0.0005	0.9985	0.0005
0.0005	0.4803	0.5187	0.0005
0.2527	0.3548	0.1713	0.2212

Download PWM File

Figure 3.4: Screenshot from the website showing an individual hCRM PWM, its sequence logo and the link to download the PWM in MEME format.

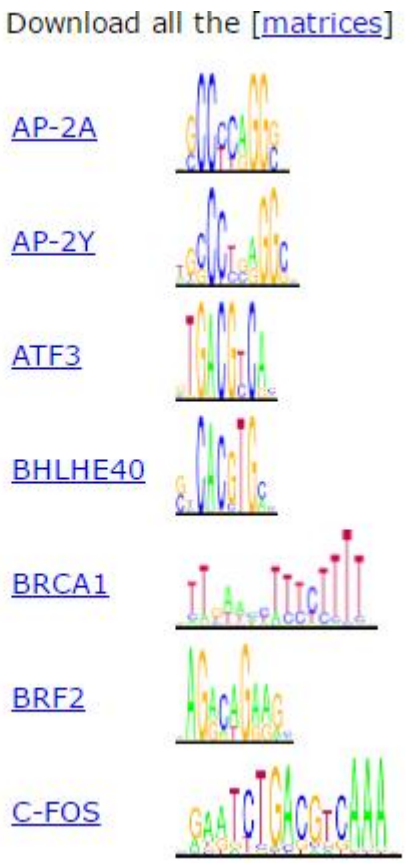


Figure 3.5: Screenshot from the website showing the hCRM PWMs and the link to bulk download them in MEME format.

3.3.4 Re-Evaluation of Pattern Matching Tools

Having shown that rGADEM generates better PWMs than other motif-discovery methods, the objective was then to reassess the performance of all the pattern matching tools investigated in the previous chapter. Performance, however, could only be assessed for the 13 transcription factors that overlap PAZAR and hCRM. The hCRM PWMs are in MEME

format which is the required format for the FIMO and MCAST pattern matching tools, and an accepted format for the matrix-scan pattern matching tool. The convert-matrix program from the RSAT suite was used to convert the PWMs into the Cluster-Buster format required for the pattern matching tools Cister, Cluster-Buster, and Comet and into the tab format required for the Patser and BayCis pattern matching tools. To convert the PWMs into the PoSSuM-PSSM format required for the PoSSuMsearch pattern matching tool, the convert-matrix program from the RSAT suite (Thomas-Chollier *et al.*, 2011) was used to convert the PWMs from MEME to TRANSFAC format and the transfac2gen program (included with the PoSSuMsearch download) was then used to convert the PWMs from TRANSFAC to the PoSSuM-PSSM format. The performance of the pattern matching tools was then reassessed by using the protocol described in section 2.2.1 with the exception that, since the evaluation was performed over the 13 transcription factors that overlapped between PAZAR and hCRM, the number of genes used in the evaluation was 167.

In the previous chapter, FIMO was identified as the best pattern matching tool for predicting individual TFBSs and MCAST was identified as the best pattern matching tool for predicting clusters of TFBSs using the JASPAR.2010 PWMs. Table 3.3 shows that these two tools still perform best using the hCRM PWMs derived here. Indeed the overall ranking of all the tools remains the same:

MCAST > Comet > Cluster-Buster > Cister > BayCis for cluster predictions and

FIMO > Patser > PoSSuMsearch > Clover > matrix-scan for individual predictions.

While evaluated on slightly different datasets, comparing the results in Table 2.3 (where tool evaluation was performed using JASPAR.2010 PWMs over 15 transcription factors that overlap between PAZAR and JASPAR) with the results in Table 3.3 (where tool evaluation was performed using hCRM PWMs over 13 transcription factors that overlap between PAZAR and hCRM) clearly shows that PWMs derived from ChIP-Seq data outperform PWMs derived from SELEX or individual promoter assays regardless of the choice of PWM scanning tool.

While it is possible that there is some inter-relationship between the choice of motif discovery method and the pattern matching tool used to search those motifs against a DNA sequence, this seems unlikely to be significant. The ranking of tool performance was the same when used with the JASPAR.2010 PWMs and the hCRM PWMs. Similarly, using FIMO, PWMs generated using rGADEM were shown to outperform those generated using MEME-ChIP.

	Sn	PPV	ACCg	FPRs
Individual				
FIMO	0.932	0.840	0.885	0.002
Patser	0.887	0.774	0.829	0.009
PoSSuMsearch	0.874	0.758	0.814	0.012
Clover	0.850	0.736	0.791	0.015
matrix-scan	0.830	0.718	0.772	0.018
Cluster				
MCAST	0.907	0.779	0.840	0.014
BayCis	0.792	0.688	0.738	0.024
Cister	0.829	0.723	0.774	0.022
Cluster-Buster	0.849	0.739	0.792	0.019
Comet	0.870	0.759	0.813	0.015

Table 3.3: Performance of the selected pattern matching tools using the hCRM PWMs derived in this work. Average sensitivities (*Sn*), Positive Predictive Value (*PPV*) and accuracy (*ACC*g) are reported together with the false positive rate using scrambled sequences (*FPR*s). Performance was evaluated across the 13 transcription factors that overlap the hCRM matrices and PAZAR.

3.4 Conclusions

In conclusion, it has been shown that PWMs derived from the ENCODE-ChIP-Seq data using rGADEM outperform those derived using other motif discovery methods. Consequently, the resulting hCRM dataset should be regarded as an enhanced addendum to resources such as JASPAR, HOCOMOCO, HOMER and CIS-BP. Clearly, as more ChIP-Seq data become available, additional PWMs will be able to be generated.

The hCRM matrices have been made publicly available for free download from http://www. bioinf.org.uk/tfbs/.

4 Utilising Transcription Factor Binding Site Prediction to Prioritize Candidate Somatic Driver SNVs

4.1 Introduction

4.1.1 Somatic SNVs in Cancer

Cancer is a leading cause of death worldwide. This is expected to increase further due to the ageing population with cancer being expected to surpass heart disease as the main killer (Wishart, 2015; Gutschner and Diederichs, 2012). Cancer is characterised as a group of abnormal cells that grow outside of the normal cell growth boundaries. This behaviour is characterised by six hallmarks which together form the fundamental principles of malignant tumour formation (Gutschner and Diederichs, 2012): (i) evasion of apoptosis (therefore resisting cell death), (ii) self-sufficiency in growth signals, (iii) insensitivity to anti-growth signals, (iv) sustained angiogenesis (which enables a consistent supply of nutrients and oxygen and the removal of carbon dioxide and waste products generated from metabolism), (v) limitless

replicative potential and (vi) tissue invasion and metastasis (Hanahan and Weinberg, 2000). This is summarised in Figure 4.1. The acquisition of these hallmarks is dependent on the accumulation of mutations. However, the rates of spontaneous mutation are very low due the ability of the genome maintenance systems to detect and repair mutations. Therefore, in order to acquire the above six cancer hallmarks, cancer cells need to increase the rate of mutation. This is done by mutating the various genes involved in the genome maintenance system e.g. TP53 (Hanahan and Weinberg, 2011).

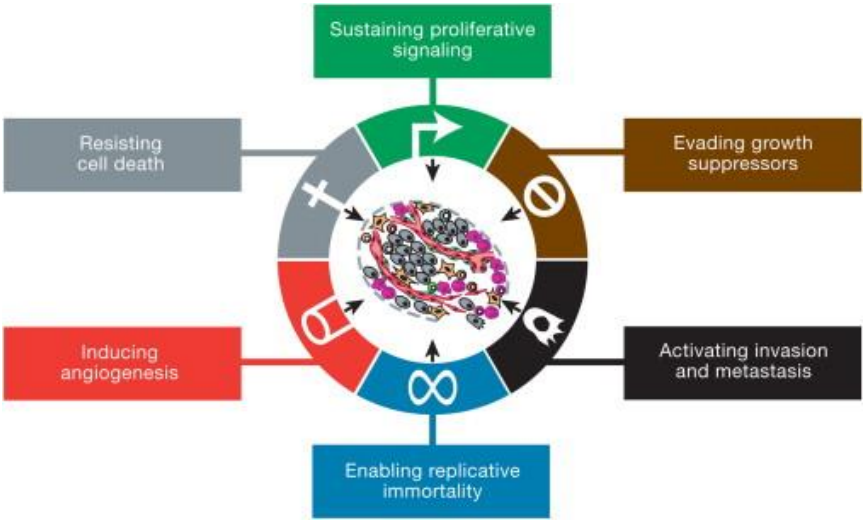


Figure 4.1: The six hallmarks of cancer (Reproduced from (Hanahan and Weinberg, 2011)).

The majority of these mutations tend to be somatic SNVs (Meyerson *et al.*, 2010). While many somatic SNVs occur in the coding regions, the majority occur in the non-coding regions (Pon and Marra, 2015). In relation to cancer, there are two types of somatic SNVs: drivers and passengers. Driver SNVs are defined as rare SNVs that confer a selective growth advantage to the cell. Passenger SNVs are defined as SNVs that do not confer a growth

advantage to the cell, and are not directly involved in cancer formation (Vogelstein *et al.*, 2013). Common SNVs (occurring in >1% of the population) are automatically classified as passenger SNVs. Therefore, passenger SNVs comprise a mixture of rare and common SNVs (Pon and Marra, 2015). Clearly research in cancer genetics is heavily focussed on driver rather than passenger SNVs (McFarland *et al.*, 2013).

There has been a comprehensive characterisation of somatic SNVs across a large number of tumour samples. This has been enabled by the recent advances in technologies for massively parallel sequencing of DNA that allow sequencing of whole exomes and genomes (Watson *et al.*, 2013). This in turn has led to large scale projects such The Cancer Genome Atlas (TCGA) (Weinstein *et al.*, 2013), and the International Cancer Genome Consortium (ICGC) (Hudson *et al.*, 2010).

Consequently, there has been a huge rise in the number of somatic driver SNVs identified, and these are deposited in the databases COSMIC (Forbes *et al.*, 2015), ICGC (Hudson *et al.*, 2010) and TCGA (Weinstein *et al.*, 2013). The data in COSMIC are a superset of ICGC and TCGA making COSMIC the largest and most comprehensive resource of somatic cancer SNVs (Forbes *et al.*, 2015; Chin *et al.*, 2011).

The rise of whole genome sequencing of human cancers has opened up the opportunity to study the large numbers of somatic SNVs that occur in non-coding regions (Poulos *et al.*, 2015a). It has been found that over 40% of somatic SNVs occur in TFBSs, and somatic SNVs are statistically enriched in TFBSs ($P < 1 \times 10^{-10}$, two-sided Fisher's exact test) (Mathelier *et*

al., 2015b; Melton *et al.*, 2015). Somatic SNVs occurring in TFBSs have also been found to have the potential to disrupt the binding of transcription factors, thereby altering the gene expression of the corresponding gene, and therefore aiding the survival and proliferation of cancer cells (Mathelier *et al.*, 2015c; Melton *et al.*, 2015; Poulos *et al.*, 2015b). It has therefore been suggested that somatic SNVs in TFBSs are a source of unidentified somatic driver SNVs (Poulos *et al.*, 2015a), and that the identification of somatic driver SNVs in TFBSs will improve diagnosis and enable more personalised therapies (Pabinger *et al.*, 2014).

4.1.2 Prioritizing Candidate Somatic Driver SNVs in TFBSs

There is now a clear need to prioritize candidate somatic driver SNVs in TFBSs for experimental validation due to the sheer volume of data being generated. In order to fulfil this need, there has to be an improvement in the computational prediction of TFBSs. This is important in order to identify the somatic SNVs that occur in TFBSs given the limited number of experimentally characterised TFBSs. In chapters 2 and 3, independent performance evaluations were carried out to identify the best performing tools which in turn will improve TFBS prediction.

Once a somatic SNV has been found to occur in a TFBS, there needs to be a way of assessing its likely effect on transcription factor binding. A somatic SNV that occurs at a position in the TFBS that is more conserved is likely to be much more disruptive to the binding of transcription factors to DNA, than a somatic SNV that occurs at a position in the TFBS with low conservation (Cline and Karchin, 2011; Gonzalez-Perez *et al.*, 2013). This is because,

more conserved positions reflect the fact that the transcription factor requires a particular nucleotide for binding. There are several measures of assessing conservation. However, only the Shannon entropy measure is applicable to nucleotide sequences.

This is because the remainder of the measures take into account the biochemical properties of amino acids (Capra and Singh, 2007). It has therefore been suggested that the utilisation of Shannon Entropy to prioritize somatic driver SNVs that occur in TFBSs might be an effective strategy (Gonzalez-Perez *et al.*, 2013; Poulos *et al.*, 2015a; Mathelier *et al.*, 2015b; Spivakov *et al.*, 2012; Cline and Karchin, 2011; Johansson and Toh, 2010).

4.1.3 Aims of Chapter

The aims of this chapter are: firstly, to exploit the analyses in the previous two chapters in order to perform a more comprehensive prediction of precise TFBSs, secondly, to perform a comprehensive analysis of the Shannon Entropy values of somatic driver and passenger SNVs that occur in TFBSs and finally to exploit this analysis to help prioritize candidate somatic driver SNVs in TFBSs for experimental validation.

4.2 Methods

All software was locally installed.

4.2.1 Prediction of TFBSs

The standard practice of predicting the TFBSs within ChIP-Seq regions when the resulting predicted TFBSs are to be used for the identification of

SNVs in TFBSs was adopted (Mathelier *et al.*, 2015c). This ensured that a very reliable set of predicted TFBSs was obtained.

ChIP-Seq peaks were downloaded from the ENCODE project (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs/) in plain text format. This is the standard dataset used for predicting TFBSs and identifying SNVs within them (Mathelier *et al.*, 2015c). As mentioned in section 3.2.1, 61 of the 90 transcription factors represented in the ENCODE project have no access restrictions and had their corresponding ChIP-Seq peaks downloaded. These transcription factors are AP-2A, AP-2Y, ATF3, BHLHE40, BRCA1, BRF2, CHD2, C-FOS, C-JUN, C-MYC, CEBPB, CTCF, E2F1, E2F4, E2F6, EBF1, ELK4, ERRA, GATA1, GATA2, GATA3, GRP20, GTF2B, HA-E2F1, HNF4A, HSF1, IRF1, IRF3, JUND, KAP1, MAFF, MAFK, MAX, NF-E2, NF-YA, NF-YB, NFKB, NRF1, POL2, PRDM1, RFX5, RPC155, SETDB1, SPT20, SREBP1, SREBP2, STAT1, STAT2, STAT3, TAL1, TBP, TCF7L2, TFIIIC-110, TR4, USF2, YY1, ZNF143, ZNF217, ZNF263, ZNF274 and ZZZ3.

These ChIP-Seq peaks were then converted to FASTA format using the Bioconductor package ChIPpeakAnno (Zhu *et al.*, 2010; Gentleman *et al.*, 2004) because the FIMO tool (Grant *et al.*, 2011), requires the input DNA sequences to be in FASTA format. Prediction of TFBSs was carried out on these sequences using FIMO which was identified as the best performing pattern matching tool in chapter 2, and the hCRM PWMs derived using rGADEM (Mercier *et al.*, 2011), which was identified as the best performing motif discovery tool in chapter 3.

The resulting predicted TFBSs were converted to BED format using PyBedTools (Dale *et al.*, 2011). As mentioned in section 2.2.1, the coordinates of the predicted TFBSs from FIMO are relative to their larger genomic fragments (i.e. are relative coordinates), while the coordinates of the ChIP-Seq peaks obtained from the ENCODE project are genomic coordinates i.e. describing their actual location in the genome. Therefore, in order to identify SNVs in TFBSs, the coordinates of the resulting predicted TFBSs were converted from relative coordinates to genomic coordinates using the convert-feature program from RSAT (Thomas-Chollier *et al.*, 2011) with output in BED format. The genomic coordinates of the ChIP-Seq peaks obtained from the ENCODE project was provided as the source of genomic coordinates to the convert-feature program. These were converted to BED format using Pybedtools (Dale *et al.*, 2011). As mentioned in section 2.2.1, the convert-feature program requires all input files in BED format. The above steps are summarised in Figure 4.2.

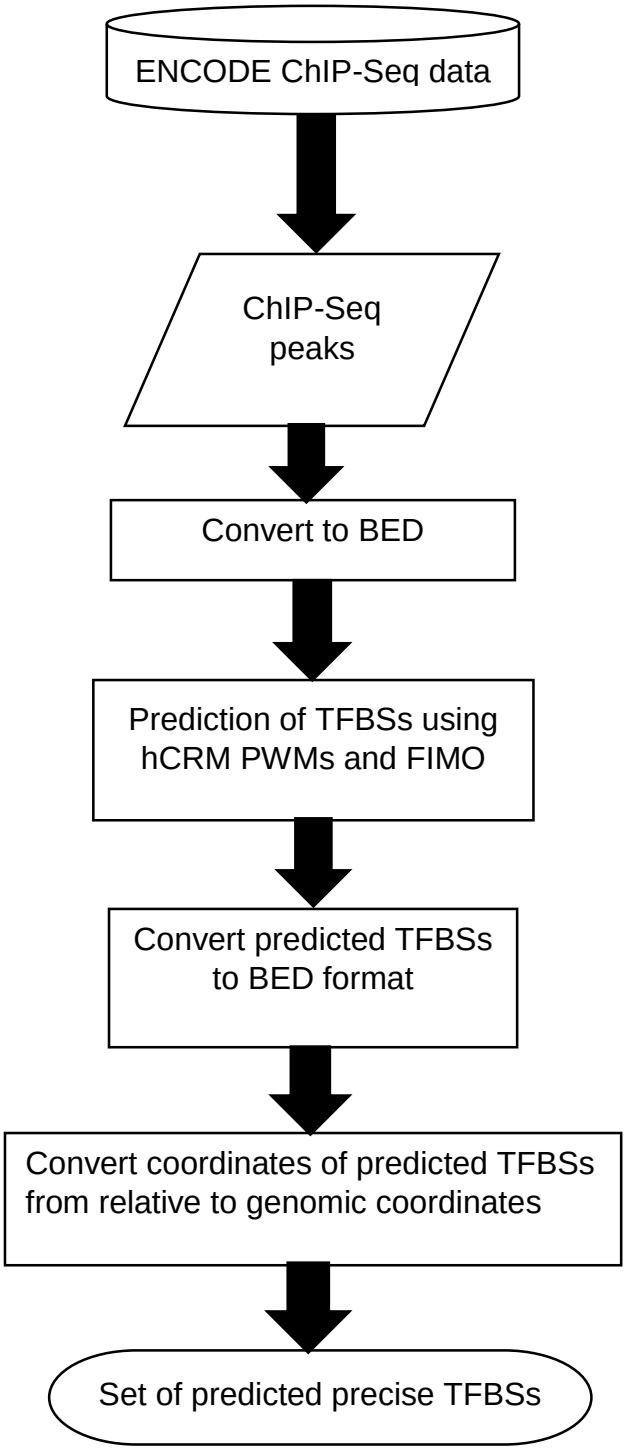


Figure 4.2: Flowchart showing the prediction of TFBSs within ENCODE ChIP-Seq peaks. See text.

4.2.2 Obtaining a Set of Somatic Cancer Driver and Passenger SNVs That Occur In TFBSs

Non-coding somatic cancer SNVs were downloaded from COSMIC (Forbes *et al.*, 2015) version 73 (the latest version at the time of doing this work) in VCF format. This VCF file is sorted by chromosome in ascending order. A total of 9 million non-coding somatic SNVs were downloaded. It was decided to use only the non-coding somatic SNVs in COSMIC given that questions have been raised whether TFBSs in protein coding regions are functional in terms of regulation of gene expression (Xing and He, 2015).

These non-coding somatic SNVs were then mapped to the predicted TFBSs and annotated with their minor allele frequency (from the 1000 genomes project). This was done using the variant_effect_predictor.pl program from the Ensembl VEP (McLaren *et al.*, 2010), which utilises the Ensembl API . Prior to this, the predicted TFBSs were sorted, compressed using bgzip (Li *et al.*, 2009b), and then indexed using Tabix (Li *et al.*, 2009b) which was required to enable the variant_effect_predictor.pl program to map the SNVs to the predicted TFBSs. The SNVs that were found to occur in the predicted TFBSs were annotated with the keyword FIMO-TFBS, together with the name of the corresponding transcription factor using the custom annotation capabilities of the Ensembl VEP. Version 81 of both the Ensembl VEP and Ensembl API was used, as these were the latest available versions at the time of doing this work. The Ensembl VEP was chosen because it has a flexible method of filtering results with the capability of the user writing their

own filter strings, and it has the ability to incorporate allele frequency information (Erzurumluoglu *et al.*, 2015).

Initially the variant_effect_predictor.pl program was very slow to run (~9 hours). In order to improve the run time, several steps were implemented as recommended on the Ensembl VEP website. These are described below:

1. First a cache file was downloaded for the human genome from ftp://ftp.ensembl.org/release81/VEP/homo_sapiens_vep_81.tar.gz
2. The --cache flag for the variant_effect_predictor.pl program was enabled in order to use the cache file. This step reduced the runtime by 3 hours.
3. The –offline flag for the variant_effect_predictor.pl program was then enabled in order to prevent the program accessing the Ensembl database, as retrieving information from only the cache file on the local file system is faster than retrieving information from the Ensembl database even if it is locally installed. This step reduced the runtime by a further 2 hours.
4. The convert_cache.pl script from the Ensembl VEP was then used to convert the cache file to a Tabix indexed file. This step reduced the runtime by a further 3 hours.
5. The Ensembl-XS package was then used to improve the run time still further as this package contains fast re-implementations in C of several key subroutines used in the variant_effect_predictor.pl program. This step reduced the runtime to 20 mins.

Therefore, the run time was greatly reduced from ~ 9 hours to 20 mins by the implementation of the above steps.

The 9 million non-coding somatic SNVs were then filtered to exclude any SNVs that were not annotated with the keyword FIMO-TFBS, and therefore, not found to occur in TFBSs. The resulting set of 159901 SNVs that were found to occur in TFBSs were then filtered further to obtain the sets of somatic driver and passenger SNVs that occur in TFBSs. The set of somatic driver SNVs in TFBSs were obtained by filtering these SNVs to exclude any SNVs that were not annotated by COSMIC as being a known driver SNV.

A total of 72329 somatic driver SNVs that occurred in TFBSs were obtained. As a set of experimentally characterised passenger SNVs is not available, the standard practice of selecting only common SNVs (with a minor allele frequency >0.01) was employed to classify these SNVs as passengers (Pon and Marra, 2015). A total of 87572 somatic passenger SNVs that occurred in TFBSs was obtained. These steps are summarised in Figure 4.3. The filtering was done using the filter_vep.pl program from the Ensembl VEP.

Both somatic driver and somatic passenger SNVs were found to occur in the set of TFBSs corresponding to transcription factors used in this work with the exception of ERRA and IRF1 where only somatic driver SNVs are found to occur.

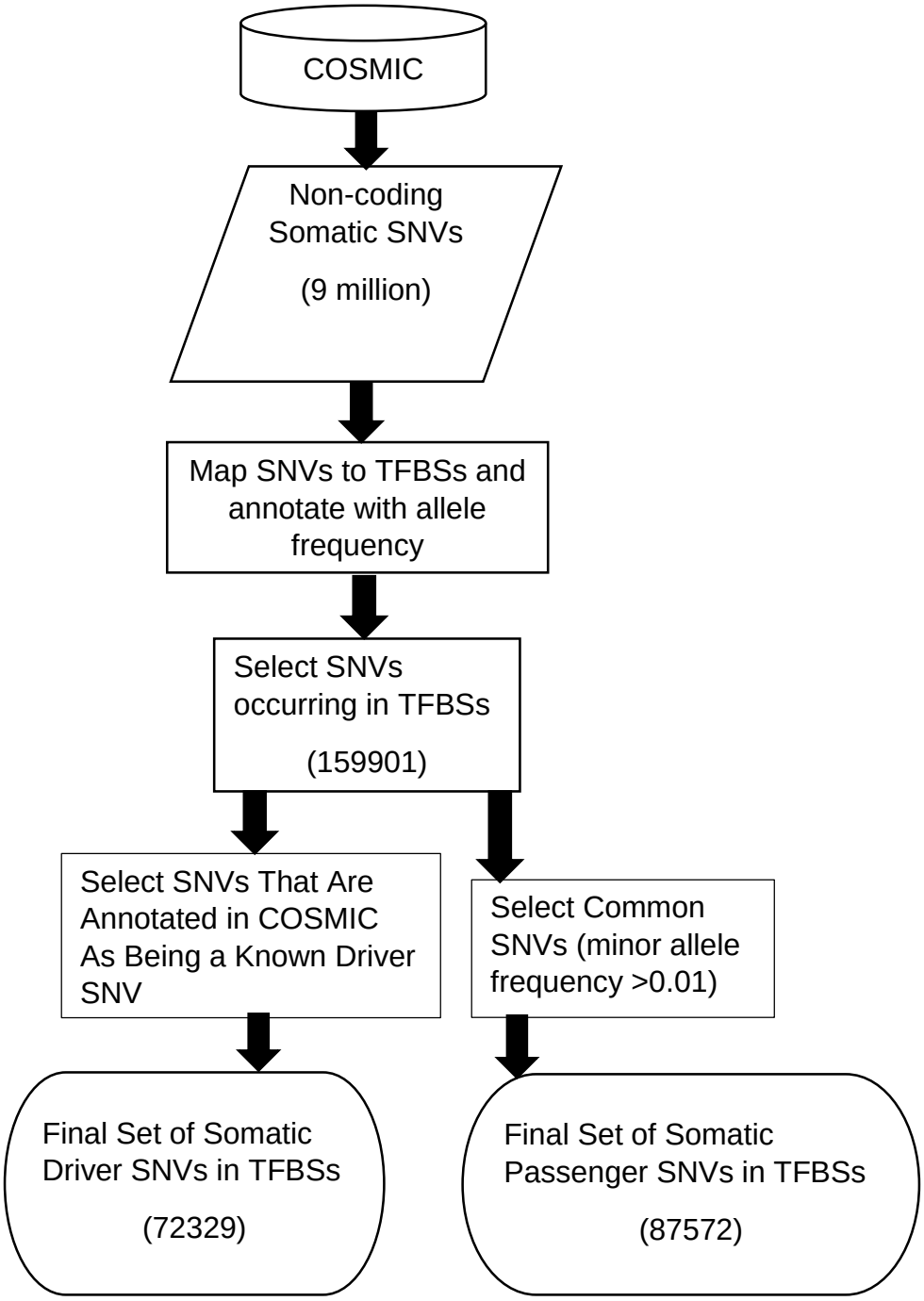


Figure 4.3: Flowchart summarising the steps taken to obtain the set of somatic driver and passenger SNVs in TFBSs. See text.

4.2.3 Calculation of Shannon Entropies for Somatic Driver and Passenger SNVs in TFBSs

Shannon Entropies were calculated for the somatic driver and passenger SNVs in TFBSs using the TFBSTools package in Bioconductor (http://www.bioconductor.org/packages/release/bioc/html/TFBSTools.html). Shannon Entropy (H) is calculated as follows

$$H = -\sum_{i} p_i \, log_b \, p_i \quad (4.1)$$

Where p_i is the probability of the character i and b is the base of the logarithm. This is normally 2 (Spivakov *et al.*, 2012). For DNA, the Shannon entropy ranges from 0 to 2 where 0 denotes complete conservation and 2 denotes an equal probability of all four bases when b= 2.

4.3 Results and Discussion

A histogram of the Shannon entropies of somatic driver and passenger SNVs in TFBSs was plotted using the ggplot2 package (Wickham, 2009) in R (Team, 2014). This is shown in Figure 4.4.

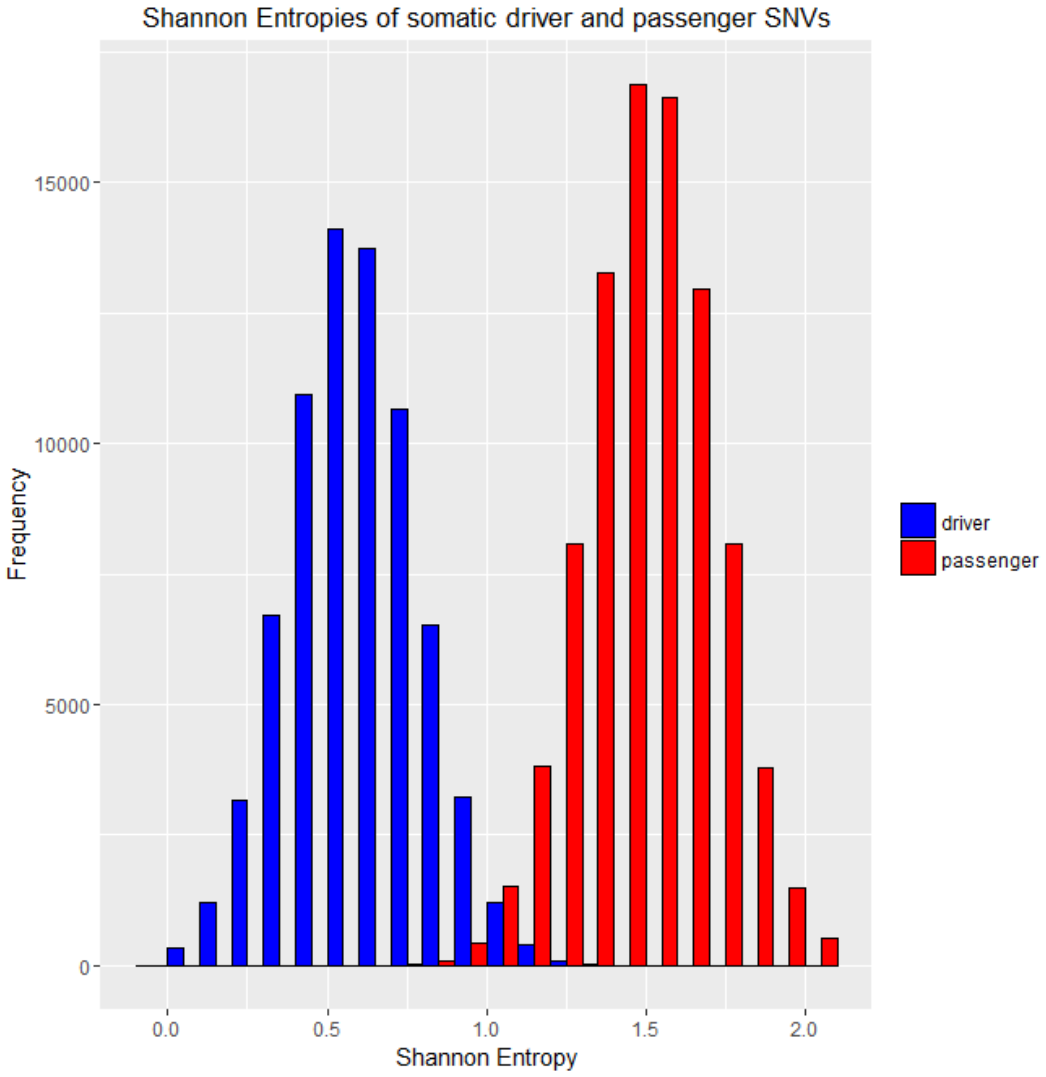


Figure 4.4: Shannon Entropies of somatic driver and passenger SNVs

Figure 4.4 shows that somatic driver SNVs in TFBSs tend to have lower Shannon entropies (i.e. be at conserved positions within TFBSs) while somatic passenger SNVs in TFBSs tend to have higher Shannon entropies (i.e. be at variable positions within TFBSs). Figure 4.4 also shows that the Shannon entropies of the somatic driver and passenger SNVs are normally distributed.

While the separation was very clear, a two sample Welch's t-test was performed in order to assess whether there was a significant difference between the means of the Shannon entropies of the somatic cancer driver SNVs in TFBSs and the Shannon entropies of the somatic cancer passenger SNVs in TFBSs. This was done using the t.test function in R (Team, 2014) which defaults to the two sample Welch t-test. This was chosen because the datasets are normally distributed, but with different variances and sample sizes.

As expected, two sample Welch t-test showed a very significant difference in the means of Shannon entropies between the somatic driver and passenger SNVs with a t statistic of -1002.561 at 154574.5 degrees of freedom and a p-value of $< 2.2 \times 10^{-16}$.

These results suggest that there are clear signals in terms of the Shannon entropies of the somatic driver and passenger SNVs in TFBSs which could be used to prioritize somatic driver SNVs in TFBSs for experimental validation.

4.3.1 Evaluating the Ability of Shannon Entropy to Prioritize Candidate Somatic Driver SNVs in TFBSs

Figure 4.4 also shows an overlap between the Shannon entropies of somatic driver and passenger SNVs in TFBSs. Therefore, in order to use Shannon entropy to prioritize candidate somatic driver SNVs in TFBSs effectively, the

optimum Shannon entropy threshold (i.e. the threshold with the best performance) within this overlap needs to be found. An SNV with a Shannon entropy below this threshold can then be considered as a candidate somatic driver SNV, while an SNV with a Shannon entropy above this threshold can be considered as a candidate somatic passenger SNV.

In order to find this optimum Shannon entropy threshold, the threshold was varied along the set of Shannon entropies that were calculated for the somatic cancer driver and passenger SNVs occurring in TFBSs. This was done in steps of 0.1 from 0 to 2 (the full range of Shannon entropies). A driver SNV with a Shannon entropy value at or below the threshold was counted as a true positive (TP), while a driver SNV with a Shannon Entropy value above the threshold was counted as a false negative (FN). In contrast, a passenger SNV with a Shannon entropy value at or below the threshold was counted as a false positive (FP), while a passenger SNV with a Shannon entropy value above the threshold was counted as a true negative (TN). This was done using R (Team, 2014). For each value of the Shannon entropy threshold, the Matthews Correlation Coefficient (MCC) (Matthews, 1975) was calculated to assess performance.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4.2)$$

The MCC was chosen as a performance indicator because, it utilizes information on true positives, false positives, true negatives and false

negatives, and therefore evaluates the performance in a more balanced manner than for example sensitivity or specificity (Baldi *et al.*, 2000). A graph of MCC against Shannon entropy threshold was then generated using R (Team, 2014) as shown in Figure 4.5.

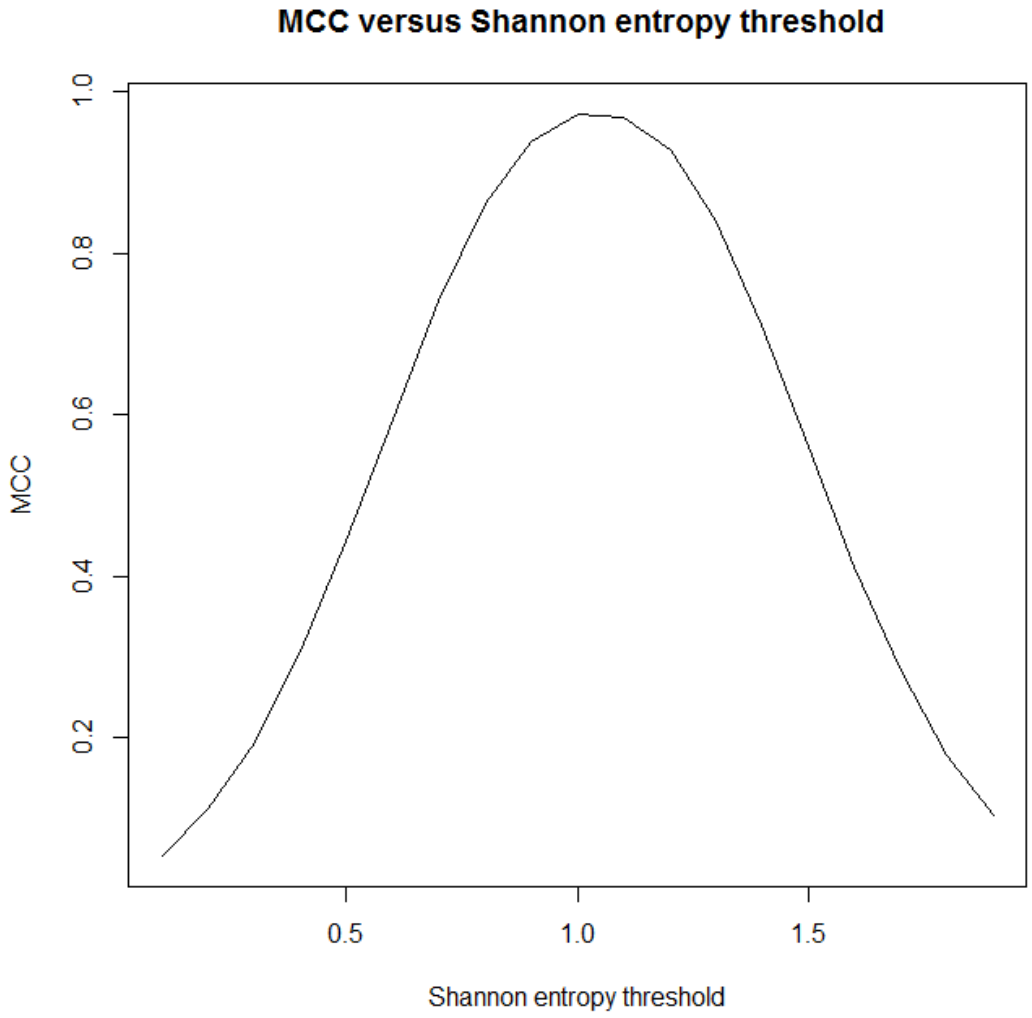


Figure 4.5: MCC plotted against Shannon entropy threshold for the full range of Shannon entropies (0 to 2).

Figure 4.5 shows that the optimum Shannon entropy threshold lies between 1 and 1.1. In order to identify the optimum Shannon entropy threshold more

precisely, the protocol used to calculate the MCC described in section 4.3.1 was repeated, but this time between 1 and 1.1 in steps of 0.01. Another graph of MCC against Shannon entropy threshold was generated using R (Team, 2014) as shown in Figure 4.6.

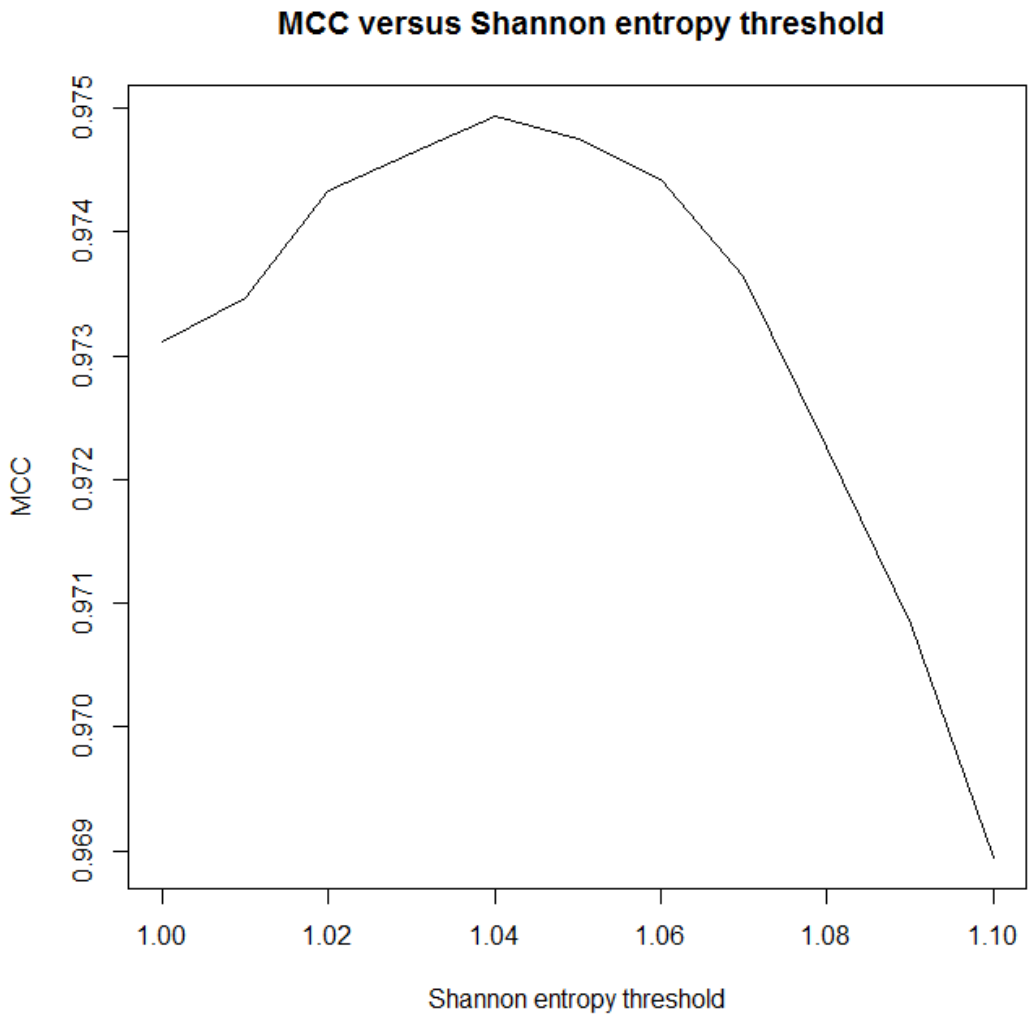


Figure 4.6: MCC plotted against Shannon entropy threshold focusing on the Shannon entropies between 1 and 1.1.

Figure 4.6 shows that the optimum Shannon entropy threshold is 1.04.Therefore an SNV with a Shannon entropy value at or below 1.04 can

be considered as a candidate somatic driver SNV, while an SNV with a Shannon entropy value above 1.04 can be considered as a candidate somatic passenger SNV.

4.4 Conclusions

The rapid growth in the whole genome sequencing of human cancers has opened up the opportunity to analyse and interpret the somatic SNVs that are present in TFBSs. Somatic driver SNVs in TFBSs tend to disrupt transcription factor binding leading to altered gene expression and consequently aiding cell proliferation and survival. Somatic passenger SNVs on the other hand do not disrupt transcription factor binding which in turn does not alter gene expression and therefore cell proliferation and survival would not be aided. Therefore there is a need to prioritize somatic driver SNVs in TFBSs.

In this work, the analysis of a set of somatic driver and passenger SNVs revealed that there were clear signals in terms of Shannon Entropy value. Somatic driver SNVs had lower Shannon entropy values (i.e. are at conserved positions within TFBSs) while somatic passenger SNVs had higher Shannon entropy values (i.e. are at variable positions within TFBSs). This was subsequently exploited to identify the optimum Shannon entropy threshold value which could be used to prioritize candidate somatic driver SNVs in TFBSs.

This work is potentially of immense value for the identification of novel somatic driver SNVs. This in turn will improve diagnosis and enable more personalised therapies.

5 Conclusions

5.1 Improving the prediction of TFBSs

The coupling of SNVs identified through whole genome sequencing with the publicly available ChIP-Seq regions that have resulted from the ENCODE project provides the opportunity to reveal novel SNVs that occur in TFBSs, which cause inherited diseases, and act as driver SNVs in cancer. In the long term, this will assist in improving the diagnosis of inherited diseases (where 30-50% of causal SNVs are missed), and will improve the diagnosis of, and aid in more personalised therapies for, cancer (Fratkin *et al.*, 2012).

The ENCODE ChIP-Seq regions are much longer than the precise binding site for a particular transcription factor, and therefore, the precise binding site still needs to be detected. This needs to be done by the computational prediction of TFBSs, as there are a limited number of experimentally characterised TFBSs available. Therefore, in order to make the most of this opportunity, there needs to be an improvement in the computational prediction of TFBSs.

There are two components to the computational prediction of TFBSs: a PWM and a pattern matching tool (Worsley-Hunt *et al.*, 2011).

PWMs are derived using motif discovery tools of which many are available. Some of these motif discovery tools have the capability to handle large volumes of data, while others do not. Motif discovery tools exist in both online and locally-installable forms. Only the locally-installable versions of the motif discovery tools that have the capacity to handle large volumes of

data can be used, given the sheer volume of ChIP-Seq data that has resulted from the ENCODE project.

As with the motif discovery tools, pattern matching tools exist in both online and locally-installable forms. However, just as with motif discovery, only the locally-installable versions can be used given the sheer volume of ChIP-Seq data that has resulted from the ENCODE project.

Several locally-installable pattern matching tools and motif discovery tools that are able to handle large volumes of data have been developed. However, to date, there has not been an independent performance evaluation of these tools.

In chapter 2, an independent evaluation of a set of open source and locally-installable pattern matching tools that predict both individual TFBSs and clusters of TFBSs was carried out. The pattern matching tools that predict individual TFBSs were found to outperform the pattern matching tools that predict clusters of TFBSs. The pattern matching tool that was found to have the best performance was FIMO (Grant *et al.*, 2011). The performance evaluation of the pattern matching tools was done before the evaluation of the motif discovery tools because, the evaluation of the performance of motif discovery methods requires a pattern matching tool to test the performance of the resulting PWMs. Therefore, a pattern matching tool must be selected for this purpose.

In chapter 3 an independent assessment of a set of open source and locally-installable motif discovery tools that are able to handle the large volumes of data that have arisen from the ENCODE project was carried out. The motif

discovery tool rGADEM (Mercier *et al.*, 2011) was found to have the best performance. A new set of PWMs were then generated using rGADEM. This new set of PWMs was named hCRM and has been made publicly available for free download (http://www.bioinf.org.uk/tfbs/). The set of pattern matching tools that were evaluated in chapter 2 were re-evaluated using the hCRM PWMs in order to check that the selection of the best pattern matching tool is not unduly influenced by the choice of PWMs. The pattern matching tool FIMO was still the best performing and the overall ranking of tools remained the same. However, the use of the hCRM PWMs (which were derived from ChIP-Seq data) to evaluate the pattern matching tools gave a better performance in comparison to evaluation of the pattern matching tools that made use of the JASPAR.2010 PWMs (which were derived from SELEX or individual promoter assays).

5.2 Application of TFBS Prediction to non-coding somatic cancer SNVs

In recent years, there has been a huge increase in the number of somatic cancer non-coding SNVs that have been identified due to the rise in whole genome sequencing of human cancers. This presents a unique opportunity to develop an approach to the prioritization of somatic non-coding cancer driver SNVs that occur in TFBSs for experimental validation (Watson *et al.*, 2013). Chapter 4 focusses on Shannon entropy to prioritize somatic non-coding cancer driver SNVs that occur in TFBSs. A more comprehensive prediction of TFBSs was first done by exploiting the analyses in chapters 2

and 3. Then an analysis of the Shannon entropy values of the driver and passenger SNVs occurring in TFBSs was performed. This analysis revealed that the driver SNVs tended to have low Shannon entropies (i.e. be at conserved positions within TFBSs) while the passenger SNVs tended to have high Shannon entropies (i.e. be at variable positions within TFBSs). This analysis was exploited to prioritize somatic driver SNVs in TFBSs by identifying the optimum Shannon entropy threshold for distinguishing between driver and passenger SNVs. The optimum threshold was found to be 1.04 but no somatic driver SNVs were identified with a Shannon entropy of >1.3 and no somatic passenger SNVs were identified with a Shannon entropy of <0.8.

5.3 Future Work

5.3.1 More Complex models

The PWM model is the most widely used model for TFBS prediction. However, the PWM model, is limited by its assumption that positions within a binding site are independent, something which is not true in all cases as it has been found that nucleotide interdependencies can exist (Nguyen and Androulakis, 2009; Hannenhalli, 2008).

Recently, more complex alternatives to the PWM model that take into account nucleotide interdependencies have been developed. These more complex alternatives are the "transcription factor flexible models" (TFFM) (Mathelier and Wasserman, 2013) and "Dinucleotide PWMs" (Kulakovskiy *et al.*, 2013a).

It has been found that more complex models do not outperform PWMs for the vast majority of transcription factors. However, for a small number of individual transcription factors e.g. REST , it has been suggested that the usage of more complex models could result in better performance (Weirauch *et al.*, 2013). Thus, in future, it may be worth evaluating PWMs, TFFMs and dinucleotide PWMs and selecting an appropriate model for each of these individual transcription factors.

5.3.2 Application of TFBS prediction to non-coding SNVs causing inherited diseases

There are currently very few examples of germline non-coding SNVs that have been found to cause inherited diseases (Heibel *et al.*, 2011; Ludlow *et al.*, 1996; Reijnen *et al.*, 1992; van Wijk *et al.*, 2003; Manco *et al.*, 2000). Whole genome sequencing is currently being carried out for many inherited diseases on a large scale (e.g. the UK100K project). As a result, a large amount of germline non-coding SNVs that cause inherited diseases is expected to become available in the next few years. Similarly to the work done in chapter 4, the predicted TFBSs can be used to identify germline non-coding inherited disease causing SNVs in TFBSs. An analysis of the Shannon entropy values of these SNVs and the Shannon entropy values of a set of common SNVs (occurring in >1% of the population) that have been obtained from dbSNP could then be carried out. The results of this analysis could then be exploited to prioritize non coding inherited disease causing SNVs in TFBSs by identifying the optimum Shannon entropy threshold for distinguishing between inherited disease causing and neutral SNVs.

References

Acharya, V. & Nagarajaram, H. A. 2012. Hansa: an automated method for discriminating disease and neutral human nsSNPs. *Human Mutation,* 33**,** 332-337.

Adey, A., Morrison, H., Asan, Xun, X., Kitzman, J., Turner, E., Stackhouse, B., MacKenzie, A., Caruccio, N., Zhang, X. & Shendure, J. 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology,* 11**,** R119.

Adli, M. & Bernstein, B. E. 2011. Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. *Nature Protocols,* 6**,** 1656-1668.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. & Sunyaev, S. R. 2010. A method and server for predicting damaging missense mutations. *Nature Methods,* 7**,** 248-249.

Al-Numair, N. S. & Martin, A. C. 2013. The SAAP pipeline and database: tools to analyze the impact and predict the pathogenicity of mutations. *BMC Genomics,* 14**,** 1-11.

Altmann, A., Weber, P., Bader, D., Preuß, M., Binder, E. B. & Müller-Myhsok, B. 2012. A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Human Genetics,* 130**,** 1-14.

Altshuler, D. M., Lander, E. S., Ambrogio, L., Bloom, T., Cibulskis, K., Fennell, T. J., Gabriel, S. B., Jaffe, D. B., Shefler, E. & Sougnez, C. L.

2010. A map of human genome variation from population scale sequencing. *Nature,* 467**,** 1061-1073.

Ao, W., Gaudet, J., Kent, W. J., Muttumu, S. & Mango, S. E. 2004. Environmentally Induced Foregut Remodeling by PHA-4/FoxA and DAF-12/NHR. *Science,* 305**,** 1743-1746.

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., Zhang, J. & Lewitter, F. 2013. Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. *PLoS Computational Biology,* 9**,** e1003326.

Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, 1994 Menlo Park, California. AAAI Press, 28-36.

Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. 2015. The MEME Suite. *Nucleic Acids Research,* 43**,** W39-W49.

Bailey, T. L. & Machanick, P. 2012. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Research,* 40**,** e128.

Bailey, T. L. & Noble, W. S. 2003. Searching for statistically significant regulatory modules. *Bioinformatics,* 19**,** ii16-ii25.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. & Nielsen, H. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics,* 16**,** 412-424.

Bao, L., Zhou, M. & Cui, Y. 2005. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Research,* 33**,** W480-W482.

Bardet, A. F., He, Q., Zeitlinger, J. & Stark, A. 2012. A computational pipeline for comparative ChIP-seq analyses. *Nature Protocols,* 7**,** 45-61.

Beckstette, M., Homann, R., Giegerich, R. & Kurtz, S. 2006. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics,* 7**,** 389-414.

Brenowitz, M., Senear, D. F. & Kingston, R. E. 2001. DNase I Footprint Analysis of Protein-DNA Binding. *Current Protocols in Molecular Biology,* 12.4**,** 12.4.1-12.4.16.

Bromberg, Y. & Rost, B. 2007. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Research,* 35**,** 3823-3835.

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

Buck, M. J. & Lieb, J. D. 2004. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics,* 83**,** 349-360.

Burrows, M. & Wheeler, D. J. 1994. A block-sorting lossless data compression algorithm. Digital Systems Research Centre Research Report.

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L. & Casadio, R. 2009. Functional annotations improve the predictive score of human

disease-related mutations in proteins. *Human Mutation,* 30**,** 1237-1244.

Capra, J. A. & Singh, M. 2007. Predicting functionally important residues from sequence conservation. *Bioinformatics,* 23**,** 1875-1882.

Capriotti, E. & Altman, R. B. 2011. Improving the prediction of disease-related variants using protein three-dimensional structure. *BMC Bioinformatics,* 12**,** S3.

Capriotti, E., Calabrese, R. & Casadio, R. 2006. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics,* 22**,** 2729-2734.

Cheng, T. M., Lu, Y.-E., Vendruscolo, M. & Blundell, T. L. 2008. Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms. *PLoS Computional Biology,* 4**,** e1000135.

Chin, L., Hahn, W. C., Getz, G. & Meyerson, M. 2011. Making sense of cancer genomic data. *Genes & Development,* 25**,** 534-555.

Chong, J., Buckingham, K., Jhangiani, S., Boehm, C., Sobreira, N., Smith, J., Harrell, T., McMillin, M., Wiszniewski, W., Gambin, T., Coban-Akdemir, Z., Doheny, K., Scott, A., Avramopoulos, D., Chakravarti, A., Hoover-Fong, J., Mathews, D., Witmer, P. D., Ling, H., Hetrick, K., Watkins, L., Patterson, K., Reinier, F., Blue, E., Muzny, D., Kircher, M., Bilguvar, K., López-Giráldez, F., Sutton, V. R., Tabor, Holly K., Leal, S., Gunel, M., Mane, S., Gibbs, R., Boerwinkle, E., Hamosh, A., Shendure, J., Lupski, J., Lifton, R., Valle, D., Nickerson, D. &

Bamshad, M. 2015. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *The American Journal of Human Genetics,* 97**,** 199-215.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X. & Ruden, D. M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly,* 6**,** 80-92.

Cline, M. S. & Karchin, R. 2011. Using bioinformatics to predict the functional impact of SNVs. *Bioinformatics,* 27**,** 441-448.

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research,* 38**,** 1767-1771.

Consortium, E. P. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57-74.

Dale, R. K., Pedersen, B. S. & Quinlan, A. R. 2011. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics,* 27**,** 3423-3424.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T. & Sherry, S. T. 2011. The variant call format and VCFtools. *Bioinformatics,* 27**,** 2156-2158.

de Vooght, K. M., van Wijk, R. & van Solinge, W. W. 2009. Management of gene promoter mutations in molecular diagnostics. *Clinical Chemistry,* 55**,** 698-708.

Djordjevic, M. 2007. SELEX experiments: new prospects, applications and data analysis in inferring regulatory pathways. *Biomolecular Engineering,* 24**,** 179-189.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A. & Huber, W. 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics,* 21**,** 3439-3440.

Durinck, S., Spellman, P. T., Birney, E. & Huber, W. 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols,* 4**,** 1184-1191.

Elnitski, L., Jin, V. X., Farnham, P. J. & Jones, S. J. M. 2006. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Research,* 16**,** 1455-1464.

Erzurumluoglu, A. M., Rodriguez, S., Shihab, H. A., Baird, D., Richardson, T. G., Day, I. N. & Gaunt, T. R. 2015. Identifying Highly Penetrant Disease Causal Mutations Using Next Generation Sequencing: Guide to Whole Process. *BioMed Research International,* 2015**,** 1-16.

Farnham, P. J. 2009. Insights from genomic profiling of transcription factors. *Nature Reviews Genetics,* 10**,** 605-616.

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

Ferragina, P. & Manzini, G. Opportunistic data structures with applications. Proceedings of the 41st Annual Symposium on Foundations of Computer Science, 2000 Redondo Beach, California. IEEE, 390-398.

Ferrer-Costa, C., Orozco, M. & De La Cruz, X. 2004. Sequence-based prediction of pathological mutations. *Proteins: Structure, Function, and Bioinformatics,* 57**,** 811-819.

Flicek, P. & Birney, E. 2009. Sense from sequence reads: methods for alignment and assembly. *Nature Methods,* 6**,** S6-S12.

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C. & Ward, S. 2015. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research,* 43**,** D805-D811.

Fratkin, E., Bercovici, S. & Stephan, D. A. 2012. The implications of ENCODE for diagnostics. *Nature Biotechnology,* 30**,** 1064-1065.

Frith, M. C., Fu, Y., Yu, L., Chen, J. F., Hansen, U. & Weng, Z. 2004a. Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Research,* 32**,** 1372-1381.

Frith, M. C., Hansen, U., Spouge, J. L. & Weng, Z. 2004b. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Research,* 32**,** 189-200.

Frith, M. C., Hansen, U. & Weng, Z. 2001. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics,* 17**,** 878-889.

Frith, M. C., Li, M. C. & Weng, Z. 2003. Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Research,* 31**,** 3666-3668.

Frith, M. C., Spouge, J. L., Hansen, U. & Weng, Z. 2002. Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Research,* 30**,** 3214-3224.

Galas, D. J. & Schmitz, A. 1978. DNAase footprinting a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Research,* 5**,** 3157-3170.

Garner, M. M. & Revzin, A. 1981. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Research,* 9**,** 3047-3060.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y. & Gentry, J. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology,* 5**,** R80.

Goecks, J., Nekrutenko, A., Taylor, J. & Team, T. G. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology,* 11**,** R86.

Gonzalez-Perez, A. & Lopez-Bigas, N. 2011. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *The American Journal of Human Genetics,* 88**,** 440-449.

Gonzalez-Perez, A., Mustonen, V., Reva, B., Ritchie, G. R. S., Creixell, P., Karchin, R., Vazquez, M., Fink, J. L., Kassahn, K. S., Pearson, J. V.,

Bader, G. D., Boutros, P. C., Muthuswamy, L., Ouellette, B. F. F., Reimand, J., Linding, R., Shibata, T., Valencia, A., Butler, A., Dronov, S., Flicek, P., Shannon, N. B., Carter, H., Ding, L., Sander, C., Stuart, J. M., Stein, L. D. & Lopez-Bigas, N. 2013. Computational approaches to identify functional genetic variants in cancer genomes. *Nature Methods,* 10**,** 723-729.

Goodstadt, L. 2010. Ruffus: a lightweight Python library for computational pipelines. *Bioinformatics,* 26**,** 2778-2779.

Grant, C. E., Bailey, T. L. & Noble, W. S. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics,* 27**,** 1017-1018.

Graur, D., Zheng, Y., Price, N., Azevedo, R. B., Zufall, R. A. & Elhaik, E. 2013. On the immortality of television sets:“function” in the human genome according to the evolution-free gospel of ENCODE. *Genome Biology and Evolution,* 5**,** 578-590.

Griffith, O. L., Montgomery, S. B., Bernier, B., Chu, B., Kasaian, K., Aerts, S., Mahony, S., Sleumer, M. C., Bilenky, M. & Haeussler, M. 2008. ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Research,* 36**,** D107-D113.

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. 2007. Quantifying similarity between motifs. *Genome Biology,* 8**,** R24.

Gutschner, T. & Diederichs, S. 2012. The hallmarks of cancer: a long non-coding RNA point of view. *RNA Biology,* 9**,** 703-19.

Hanahan, D. & Weinberg, R. A. 2000. The hallmarks of cancer. *Cell,* 100**,** 57-70.

Hanahan, D. & Weinberg, R. A. 2011. Hallmarks of Cancer: The Next Generation. *Cell,* 144**,** 646-674.

Hannenhalli, S. 2008. Eukaryotic transcription factor binding sites—modeling and integrative search methods. *Bioinformatics,* 24**,** 1325-1331.

Hayden, E. C. 2014. The $1,000 genome. *Nature,* 507**,** 294-295.

Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R. & Ordoukhanian, P. 2014. Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques,* 56**,** 61-77.

Heibel, S. K., Ah Mew, N., Caldovic, L., Daikhin, Y., Yudkoff, M. & Tuchman, M. 2011. N-carbamylglutamate enhancement of ureagenesis leads to discovery of a novel deleterious mutation in a newly defined enhancer of the NAGS gene and to effective therapy. *Human mutation,* 32**,** 1153-1160.

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

Hellman, L. M. & Fried, M. G. 2007. Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions. *Nature Protocols,* 2**,** 1849-1861.

Hertz, G. Z. & Stormo, G. D. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics,* 15**,** 563-577.

Hu, M., Yu, J., Taylor, J. M. G., Chinnaiyan, A. M. & Qin, Z. S. 2010. On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Research,* 38**,** 2154-2167.

Hudson, T. J., Anderson, W., Aretz, A., Barker, A. D., Bell, C., Bernabé, R. R., Bhan, M., Calvo, F., Eerola, I. & Gerhard, D. S. 2010. International network of cancer genome projects. *Nature,* 464**,** 993-998.

Hughes, J. D., Estep, P. W., Tavazoie, S. & Church, G. M. 2000. Computational identification of Cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae1. *Journal of Molecular Biology,* 296**,** 1205-1214.

Hunt, R. W., Mathelier, A., Del Peso, L. & Wasserman, W. W. 2014. Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment. *BMC Genomics,* 15**,** 472-491.

Jamuar, S. S. & Tan, E.-C. 2015. Clinical application of next-generation sequencing for Mendelian diseases. *Human Genomics,* 9**,** 10-16.

Johansson, F. & Toh, H. 2010. A comparative study of conservation and variation scores. *BMC Bioinformatics,* 11**,** 388-399.

Joshua, H., Peter, K., Nicolas, N. & Peter, P. 2011. ChIP-chip versus ChIP-seq: Lessons for experimental design and data analysis. *BMC Genomics,* 12**,** 134-146.

Jothi, R., Cuddapah, S., Barski, A., Cui, K. & Zhao, K. 2008. Genome-wide identification of in vivo protein–DNA binding sites from ChIP-Seq data. *Nucleic Acids Research,* 36**,** 5221-5231.

Kärkkäinen, J. 2007. Fast BWT in small space by blockwise suffix sorting. *Theoretical Computer Science,* 387**,** 249-257.

Kel, A. E., Gößling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V. & Wingender, E. 2003. MATCH(TM): a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research,* 31**,** 3576-3579.

Khan, S. & Vihinen, M. 2007. Spectrum of disease-causing mutations in protein secondary structures. *BMC Structural Biology,* 7**,** 56-74.

Klauer, A. A. & van Hoof, A. 2012. Degradation of mRNAs that lack a stop codon: a decade of nonstop progress. *Wiley Interdisciplinary Reviews: RNA,* 3**,** 649-660.

Klug, W., S., Cummings, M., R., Spencer, C., A. & Palladino, M., A. 2012. *Concepts of genetics* Boston,Massachusetts: , Pearson.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L. & Wilson, R. K. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research,* 22**,** 568-576.

Kulakovskiy, I., Boeva, V., Favorov, A. & Makeev, V. 2010. Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics,* 26**,** 2622-2623.

Kulakovskiy, I., Levitsky, V., Oshchepkov, D., Bryzgalov, L., Vorontsov, I. & Makeev, V. 2013a. From binding motifs in ChIP-seq data to improved models of transcription factor binding sites. *Journal of Bioinformatics and Computational biology,* 11**,** 1-12.

Kulakovskiy, I. V., Medvedeva, Y. A., Schaefer, U., Kasianov, A. S., Vorontsov, I. E., Bajic, V. B. & Makeev, V. J. 2013b. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Research,* 41**,** D195-D202.

Kumar, P., Henikoff, S. & Ng, P. C. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols,* 4**,** 1073-1081.

Kurmangaliyev, Y. Z., Sutormin, R. A., Naumenko, S. A., Bazykin, G. A. & Gelfand, M. S. 2013. Functional implications of splicing polymorphisms in the human genome. *Human Molecular Genetics,* 22**,** 3449-3459.

Langmead, B. & Salzberg, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods,* 9**,** 357-359.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology,* 10**,** R25.

Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., Ley, T. J., Mardis, E. R., Wilson, R. K. & Ding, L. 2012. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics,* 28**,** 311-317.

Li, B., Krishnan, V. G., Mort, M. E., Xin, F., Kamati, K. K., Cooper, D. N., Mooney, S. D. & Radivojac, P. 2009a. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics,* 25**,** 2744-2750.

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics,* 25**,** 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. 2009b. The sequence alignment/map format and SAMtools. *Bioinformatics,* 25**,** 2078-2079.

Li, H., Ruan, J. & Durbin, R. 2008a. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research,* 18**,** 1851-1858.

Li, R., Li, Y., Kristiansen, K. & Wang, J. 2008b. SOAP: short oligonucleotide alignment program. *Bioinformatics,* 24**,** 713-714.

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K. & Wang, J. 2009c. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics,* 25**,** 1966-1967.

Lin, H., Zhang, Z., Zhang, M. Q., Ma, B. & Li, M. 2008a. ZOOM! Zillions of oligos mapped. *Bioinformatics,* 24**,** 2431-2437.

Lin, T., Ray, P., Sandve, G. K., Uguroglu, S. & Xing, E. P. 2008b. Baycis: a bayesian hierarchical hmm for cis-regulatory module decoding in metazoan genomes. *Lecture Notes in Computer Science,* 4955**,** 66-81.

Lopes, M. C., Joyce, C., Ritchie, G. R., John, S. L., Cunningham, F., Asimit, J. & Zeggini, E. 2012. A combined functional annotation score for non-synonymous variants. *Human Heredity,* 73**,** 47-51.

Ludlow, L. B., Schick, B. P., Budarf, M. L., Driscoll, D. A., Zackai, E. H., Cohen, A. & Konkle, B. A. 1996. Identification of a mutation in a GATA binding site of the platelet glycoprotein Ibβ promoter resulting

in the Bernard-Soulier syndrome. *Journal of Biological Chemistry,* 271**,** 22076-22080.

Ma, W., Noble, W. S. & Bailey, T. L. 2014. Motif-based analysis of large nucleotide data sets using MEME-ChIP. *Nature Protocols,* 9**,** 1428-1450.

Machanick, P. & Bailey, T. L. 2011. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics,* 27**,** 1696-1697.

Makrythanasis, P. & Antonarakis, S. E. 2011. From sequence to functional understanding: the difficult road ahead. *Genome Medicine,* 3**,** 21-24.

Manco, L., Ribeiro, M. L., Maximo, V., Almeida, H., Costa, A., Freitas, O., Barbot, J., Abade, A. & Tamagnini, G. 2000. A new PKLR gene mutation in the R-type promoter region affects the gene transcription causing pyruvate kinase deficiency. *British Journal of Haematology,* 110**,** 993-997.

Maquat, L. E. 2004. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nature Reviews Molecular Cell Biology,* 5**,** 89-99.

Maquat, L. E. 2005. Nonsense-mediated mRNA decay in mammals. *Journal of Cell Science,* 118**,** 1773-1776.

Marine, R., Polson, S., Ravel, J., Hatfull, G., Russell, D., Sullivan, M., Syed, F., Dumas, M. & Wommack, K. E. 2011. Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. *Applied and Environmental Microbiology,* 77**,** 8071-8079.

Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C.-y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G. & Worsley-Hunt, R. 2015a. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research,* 44**,** D110-D115.

Mathelier, A., Lefebvre, C., Zhang, A. W., Arenillas, D. J., Ding, J., Wasserman, W. W. & Shah, S. P. 2015b. Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. *Genome Biology,* 16**,** 84-101.

Mathelier, A., Shi, W. & Wasserman, W. W. 2015c. Identification of altered cis-regulatory elements in human disease. *Trends in Genetics,* 31**,** 67-76.

Mathelier, A. & Wasserman, W. W. 2013. The next generation of transcription factor binding site prediction. *PLoS Computational Biology,* 9**,** e1003214.

Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Chen, C.-y., Chou, A. & Ienasescu, H. 2013. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research,* 42**,** D142-D147.

Matthews, B. W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta-Protein Structure,* 405**,** 442-451.

Matys, V., Kel-Margoulis, O., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M. & Hornischer, K. 2006.

TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic acids research,* 34**,** D108-D110.

McFarland, C. D., Korolev, K. S., Kryukov, G. V., Sunyaev, S. R. & Mirny, L. A. 2013. Impact of deleterious passenger mutations on cancer progression. *Proceedings of the National Academy of Sciences,* 110**,** 2910-2915.

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

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P. & Cunningham, F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics,* 26**,** 2069-2070.

Melton, C., Reuter, J. A., Spacek, D. V. & Snyder, M. 2015. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nature Genetics,* 47**,** 710-716.

Mercier, E., Droit, A., Li, L., Robertson, G., Zhang, X. & Gottardo, R. 2011. An Integrated Pipeline for the Genome-Wide Analysis of Transcription Factor Binding Sites from ChIP-Seq. *PLoS One,* 6**,** e16432.

Metzker, M. L. 2009. Sequencing technologies—the next generation. *Nature Reviews Genetics,* 11**,** 31-46.

Meyerson, M., Gabriel, S. & Getz, G. 2010. Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics,* 11**,** 685-696.

Moore, G. E. 1998. Cramming more components onto integrated circuits. *Proceedings of the IEEE,* 86**,** 82-85.

Narlikar, L. & Ovcharenko, I. 2009. Identifying regulatory elements in eukaryotic genomes. *Briefings in Functional Genomics & Proteomics,* 8**,** 215-230.

Natrajan, R. & Reis-Filho, J. S. 2011. Next-generation sequencing applied to molecular diagnostics. *Expert Review of Molecular Diagnostics,* 11**,** 425-444.

Nguyen, T. T. & Androulakis, I. P. 2009. Recent advances in the computational discovery of transcription factor binding sites. *Algorithms,* 2**,** 582-605.

Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. 2011. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics,* 12**,** 443-451.

Orgel, L. E. & Crick, F. H. 1980. Selfish DNA: the ultimate parasite. *Nature,* 284**,** 604-607.

Ovaska, K., Laakso, M., Haapa-Paananen, S., Louhimo, R., Chen, P., Aittomäki, V., Valo, E., Núñez-Fontarnau, J., Rantanen, V. & Karinen, S. 2010. Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Medicine,* 2**,** 65-77.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J. & Trajanoski, Z. 2014. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics,* 15**,** 256-278.

Park, P. J. 2009. ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics,* 10**,** 669-680.

Pon, J. R. & Marra, M. A. 2015. Driver and passenger mutations in cancer. *The Annual Review of Pathology:Mechanisms of Disease,* 10**,** 25-50.

Portales-Casamar, E., Arenillas, D., Lim, J., Swanson, M. I., Jiang, S., McCallum, A., Kirov, S. & Wasserman, W. W. 2009. The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Research,* 37**,** D54-D60.

Portales-Casamar, E., Thongjuea, S., Kwon, A. T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W. W. & Sandelin, A. 2010. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Research,* 38**,** D105-D110.

Poulos, R. C., Sloane, M. A., Hesson, L. B. & Wong, J. W. H. 2015a. The search for cis -regulatory driver mutations in cancer genomes. *Oncotarget,* 6**,** 32509-32525.

Poulos, R. C., Thoms, J., Shah, A., Beck, D., Pimanda, J. E. & Wong, J. W. H. 2015b. Systematic Screening of Promoter Regions Pinpoints Functional Cis-Regulatory Mutations in a Cutaneous Melanoma Genome. *Molecular Cancer Research,* 13**,** 1218-1226.

Quinlan, A. R. & Hall, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* 26**,** 841-842.

Read, A. & Donnai, D. 2011. *New clinical genetics* Banbury, Scion Publishing Ltd.

Reijnen, M. J., Sladek, F. M., Bertina, R. M. & Reitsma, P. H. 1992. Disruption of a binding site for hepatocyte nuclear factor 4 results in hemophilia B Leyden. *Proceedings of the National Academy of Sciences,* 89**,** 6300-6303.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N. & Kanin, E. 2000. Genome-wide location and function of DNA binding proteins. *Science,* 290**,** 2306-2309.

Reva, B., Antipin, Y. & Sander, C. 2011. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids Research,* 39**,** e118.

Rice, P., Longden, I. & Bleasby, A. 2000. EMBOSS: the European molecular biology open software suite. *Trends in Genetics,* 16**,** 276-277.

Rumble, S. M., Lacroute, P., Dalca, A. V., Fiume, M., Sidow, A. & Brudno, M. 2009. SHRiMP: accurate mapping of short color-space reads. *PLoS Computational Biology,* 5**,** e1000386.

Sadedin, S. P., Pope, B. & Oshlack, A. 2012. Bpipe: a tool for running and managing bioinformatics pipelines. *Bioinformatics,* 28**,** 1525-1526.

Sand, O., Turatsinze, J.-V. & van Helden, J. 2008. Evaluating the prediction of cis-acting regulatory elements in genome sequences. *In:* Frishman, D. & Valencia, A. (eds.) *Modern Genome Annotation.* Springer.

Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W. & Lenhard, B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research,* 32**,** D91-D94.

Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K. & Gerstein, M. B. 2011. The real cost of sequencing: higher than you think! *Genome Biology,* 12**,** 125.

Schnekenberg, R. P. & Németh, A. H. 2013. Next-generation sequencing in childhood disorders. *Archives of disease in childhood,* 99**,** 284-290.

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., Day, I. N. & Gaunt, T. R. 2013. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human Mutation,* 34**,** 57-65.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G. & Kasprzyk, A. 2009. BioMart–biological queries made easy. *BMC Genomics,* 10**,** 22-34.

Spivakov, M., Akhtar, J., Kheradpour, P., Beal, K., Girardot, C., Koscielny, G., Herrero, J., Kellis, M., Furlong, E. E. & Birney, E. 2012. Analysis of variation at transcription factor binding sites in Drosophila and humans. *Genome Biology,* 13**,** R49.

Stone, E. A. & Sidow, A. 2005. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Research,* 15**,** 978-986.

Team, R. C. 2014. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P. & Moreau, Y. 2001. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics,* 17**,** 1113-1122.

Thomas-Chollier, M., Defrance, M., Medina-Rivera, A., Sand, O., Herrmann, C., Thieffry, D. & van Helden, J. 2011. RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Research,* 39**,** W86-W91.

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A. & Narechania, A. 2003. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Research,* 13**,** 2129-2141.

Tian, J., Wu, N., Guo, X., Guo, J., Zhang, J. & Fan, Y. 2007. Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines. *BMC Bioinformatics,* 8**,** 450-459.

Tuerk, C. & Gold, L. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science,* 249**,** 505-510.

Turatsinze, J. V., Thomas-Chollier, M., Defrance, M. & Van Helden, J. 2008. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature Protocols,* 3**,** 1578-1588.

Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M. & Sidow, A. 2008. Genome-wide analysis

of transcription factor binding sites based on ChIP-Seq data. *Nature Methods,* 5**,** 829-834.

van Wijk, R., van Solinge, W., Nerlov, C., Beutler, E., Gelbart, T., Rijksen, G. & Nielsen, F. 2003. Disruption of a novel regulatory element in the erythroid-specific promoter of the human PKLR gene causes severe pyruvate kinase deficiency. *Blood,* 101**,** 1596-1602.

Vlieghe, D., Sandelin, A., De Bleser, P. J., Vleminckx, K., Wasserman, W. W., Van Roy, F. & Lenhard, B. 2006. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Research,* 34**,** D95-D97.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A. & Kinzler, K. W. 2013. Cancer Genome Landscapes. *Science* 339**,** 1546-1558.

Wang, K., Li, M. & Hakonarson, H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research,* 38**,** e164.

Watson, I. R., Takahashi, K., Futreal, P. A. & Chin, L. 2013. Emerging patterns of somatic mutations in cancer. *Nature Reviews Genetics,* 14**,** 703-718.

Wei, Z., Wang, W., Hu, P., Lyon, G. J. & Hakonarson, H. 2011. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Research,* 39**,** e132.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M. & Network,

C. G. A. R. 2013. The cancer genome atlas pan-cancer analysis project. *Nature Genetics,* 45**,** 1113-1120.

Weirauch, M. T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T. R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A. & Talukder, S. 2013. Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnology,* 31**,** 126-134.

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. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell,* 158**,** 1431-1443.

Wickham, H. 2009. *ggplot2: elegant graphics for data analysis*, Springer Science & Business Media.

Wilbanks, E. G. & Facciotti, M. T. 2010. Evaluation of Algorithm Performance in ChIP-Seq Peak Detection. *PLoS One,* 5**,** e11471.

Wishart, D. S. 2015. Is Cancer a Genetic Disease or a Metabolic Disease? *EBioMedicine,* 2**,** 478-479.

Wong, K.-C. & Zhang, Z. 2014. SNPdryad: predicting deleterious non-synonymous human SNPs using only orthologous protein sequences. *Bioinformatics,* 30**,** 1112-1119.

Worsley-Hunt, R., Bernard, V. & Wasserman, W. W. 2011. Identification of cis-regulatory sequence variations in individual genome sequences. *Genome Medicine,* 3**,** 1-14.

Xing, K. & He, X. 2015. Reassessing the “Duon” Hypothesis of Protein Evolution. *Molecular Biology and Evolution,* 32**,** 1056-1062.

Yang, V. W. 1998. Eukaryotic transcription factors: identification, characterization and functions. *The Journal of Nutrition,* 128**,** 2045-2051.

Yates, C. M., Filippis, I., Kelley, L. A. & Sternberg, M. J. 2014. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *Journal of Molecular Biology,* 426**,** 2692-2701.

Yue, P., Melamud, E. & Moult, J. 2006. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics,* 7**,** 166-181.

Zambelli, F., Pesole, G. & Pavesi, G. 2012. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in Bioinformatics,* 14**,** 225-237.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M. & Li, W. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biology,* 9**,** R137.

Zhu, L. J., Gazin, C., Lawson, N. D., Pagès, H., Lin, S. M., Lapointe, D. S. & Green, M. R. 2010. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics,* 11**,** 237-247.