OÉ Gaillimh
NUI Galway

Machine Learning and High-Performance Computing: Infrastructure and Algorithms for the Genome-Scale Study of Genetic and Epigenetic Regulatory Mechanisms With Applications in Neuroscience

A thesis submitted

by

Pilib Ó Broin

to the

School of Mathematics, Statistics and Applied Mathematics,
College of Science,
National University *of* Ireland, Galway

in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

June 2014

Thesis Supervisor: Dr. Aaron Golden
Director of Research: Prof. Terry Smith

Contents

List of Figures

List of Tables

Acknowledgements

This work would not have been possible without the support and generosity of a number of people, I would like to take this opportunity to thank them for their contributions.

Firstly, I would like to thank Prof. Terry Smith, my Director of Research. I also wish to express my sincerest gratitude to Dr. Aaron Golden, without whose guidance this work would never have come to fruition. His insights, advice, and encouragement over the years have been invaluable and his constant willingness to allow me to explore my own research ideas has always been greatly appreciated.

This research is firmly rooted in the interactions we have had with some of our many basic research and clinical collaborators and I gratefully acknowledge their contributions.

I would firstly like to thank Prof. John Greally and the members of his lab for hosting my initial stay at Einstein. It was there that WASP was developed as part of the Computational Epigenomics Group. My sincere thanks to the other members of the initial development team – Andrew McLellan, Robert Dubin, and A.J. Jing – as well as to the members of the Computational Genomics Group, particularly Brent Calder and David Moskowitz. I couldn't ask for a better group with which to spend as many hours poring over design decisions, basecode, and error logs.

Also at Einstein, I would like to thank the lab of Dr. Noboru Hiroi for the 22q11.DS ChIP-seq data and for providing me with the interesting challenge of 'talking to mice'.

My thanks also to our collaborators in the labs of Prof. Jonathan Licht at Northwestern University and Dr. Anton Krumm at the University of Washington for the many fruitful interactions and for generating the ChIP-chip datasets used to test the ChIPSOM algorithm.

On a personal level I would like to express my deep and abiding gratitude to my family for their continued support in all that I do. Finally, to my wife Muireann, for her unwavering encouragement, understanding, and patience throughout the years – ní fhéadfainn é seo a dhéanamh gan thú, is tú is ansa liom.

Publications

Pilib Ó Broin, Terry Smith, Aaron Golden, **Non-tree-based alignment-free clustering of position weight matrices.** *Submitted to BMC Bioinformatics.*

Pilib Ó Broin, Bhavapriya Vaitheesvaran, Subhrajit Saha, Kirsten Hartil, Emily I Chen, Devorah Goldman, William H Fleming, Irwin J Kurland, Aaron Golden, Chandan Guha, **Intestinal microbiota derived metabolomic blood plasma markers for prior Radiation injury.** *Submitted to International Journal of Radiation Oncology*Biology*Physics.*

Pilib Ó Broin, Terry Smith, Aaron Golden, **Mumbles: A web-based entropy tool for structure determination in mouse vocalization sequences.** *In preparation.*

Behnam Nabet, Pilib Ó Broin, Kevin Shieh, Charles Y. Lin, Christine M. Will, Relja Popovic, Teresa Ezponda, James E. Bradner, Aaron A. Golden, Jonathan D. Licht, **Oncogenic receptor tyrosine kinase signaling deregulates the enhancer landscape.** *In preparation.*

Tomohisa Takahashi, Shota Okabe, Pilib Ó Broin, Akira Nishi, Takeshi Izumi, Michael Beckert, Gina Kang, Seiji Ishikawa, Ichiro Tateya, Norio Yamamoto, Jose L. Pena, Kenji Tanigaki, Aaron Golden, Takefumi Kikusui, Noboru Hiroi, **Structure and function of social communication in a genetic mouse model of developmental neuropsychiatric disorders.** *In preparation.*

Shuken Boku, Takeshi Hiramoto, Akitoyo Hishimoto, Gina Kang, Tatyana V. Michurina, Grigori Enikolopov, Pilib Ó Broin, Aaron Golden, Kenji Tanigaki, Noboru Hiroi, **Tbx1 transcriptionally regulates Pten and postnatal neurogenesis in a genetic mouse model of developmental neuropsychiatric disorders.** *In preparation.*

Andrew S McLellan, Robert A Dubin, Qiang Jing, Pilib Ó Broin, David Moskowitz, Masako Suzuki, R Brent Calder, Joseph Hargitai, Aaron Golden, John M Greally, **The Wasp System: An open source environment for managing and analyzing genomic data.** *Genomics. 2012 Dec;100(6):345-51.*

Aaron Golden, Andrew S McLellan, Robert A Dubin, Qiang Jing, Pilib Ó Broin, David Moskowitz, Zhengdong Zhang, Masako Suzuki, Joseph Hargitai, R Brent Calder, John M Greally, **The Einstein Genome Gateway using WASP - a high throughput multi-layered life sciences portal for XSEDE.** *Stud Health Technol Inform. 2012;175:182-91.*

Brandon J Thomas, Eric D Rubio, Niklas Krumm, Pilib Ó Broin, Karol Bomsztyk, Piri Welcsh, John M Greally, Aaron A Golden, Anton Krumm, **Allele-specific transcriptional elongation regulates monoallelic expression of the IGF2BP1 gene.** *Epigenetics Chromatin. 2011 Aug 3;4:14.*

Andrew S McLellan, Robert A Dubin, Qiang Jing, Shahina B Maqbool, Raul Olea, Gael Westby, Pilib Ó Broin, Melissa J Fazzari, Deyou Zheng, Masako Suzuki, John M Greally, **The Einstein Center for Epigenomics: studying the role of epigenomic dysregulation in human disease.** *Epigenomics. 2009 Oct;1(1):33-8.*

Marianne K-H Kim, Thomas J McGarry, Pilib Ó Broin, Jared M Flatow, Aaron A-J Golden, Jonathan D Licht, **An integrated genome screen identifies the Wnt signaling pathway as a major target of WT1.** *Proc Natl Acad Sci U S A. 2009 Jul 7;106(27):11154-9.*

Abstract

The advent of next-generation sequencing (NGS) has fundamentally changed modern genomics research. These sequencers generate terabytes of data and necessitate the use, not only of high-performance compute (HPC) clusters for data processing and storage, but also of intelligent, scalable algorithms for pattern discovery and data mining. This thesis details the development of infrastructure and algorithms which automate much of this data analysis process allowing bench biologists to remain focused on the scientific questions that drive them, rather than the informatics challenges associated with these new platforms. We describe WASP, one of the first end-to-end systems to handle all aspects of NGS data generation, including sample submission, laboratory information management system (LIMS) functionality, and assay-specific processing pipelines. Furthermore, we present two machine learning algorithms for the secondary analysis of ChIP-seq data, the first, based on the use of self-organising maps (SOMs) for improved *de novo* motif discovery, and the second, which uses genetic algorithms (GAs) to automatically cluster transcription factor binding motifs. Finally, we present an application of this infrastructure and these techniques to the study of the role of the TBX1 transcription factor in 22q11.2 Deletion Syndrome, examining its putative role in neural development, adult neurogenesis, autism spectrum disorder (ASD), and schizophrenia.

CHAPTER 1

Introduction

1.1 The Digital Genome and the Data Deluge

Since 1995 and the whole-genome shotgun sequencing (WGSS) [1] of the bacterial genome *Haemophilus influenzae* [2], the complete genomes of over four thousand different organisms (including, in 2001, the human genome [3, 4]) have been sequenced. There are also currently over 33,500 genome sequencing projects listed at the Genomes Online Database[1] (GOLD) [5] – this represents a vast quantity of data, the volume of which has grown exponentially in the past few years with the increasing adoption of paradigm-shifting massively-parallel sequencing (MPS) technologies or next-generation sequencers (NGS). These new sequencers operate by generating millions of short genomic reads in parallel, greatly reducing the cost (Figure 1.1) and time associated with sequencing [6]. They also however create some interesting challenges for existing computational infrastructure, given that a single experiment can take days to run and result in terabytes of raw data. The possibilities for medical science to use all of this data to better understand disease and provide more sophisticated, perhaps even personalised treatment, depends therefore on the ability to mine this information in useful ways. This requires making use of advanced statistical techniques and drawing on expertise from the fields of artificial intelligence (AI) and machine learning (ML). Techniques from these fields (such as image analysis, pattern recognition, search and optimisation, and probabilistic reasoning) readily lend themselves to the analysis of large, complex datasets, where researchers are often led to new insights or new lines of inquiry by previously unlooked-for patterns revealed in the data.

One of the key ways in which this new sequence data is used is in the genome-scale study of gene regulation. Next-generation sequencers can be used to map, for the the first time in a high-resolution manner, the genome-wide locations of sites where regulatory mechanisms can act to influence the expression of a gene or group of genes [7, 8, 9]. In the next section we introduce some of those regulatory mechanisms and outline their importance in practically all aspects of cellular

[1] http://www.genomesonline.org

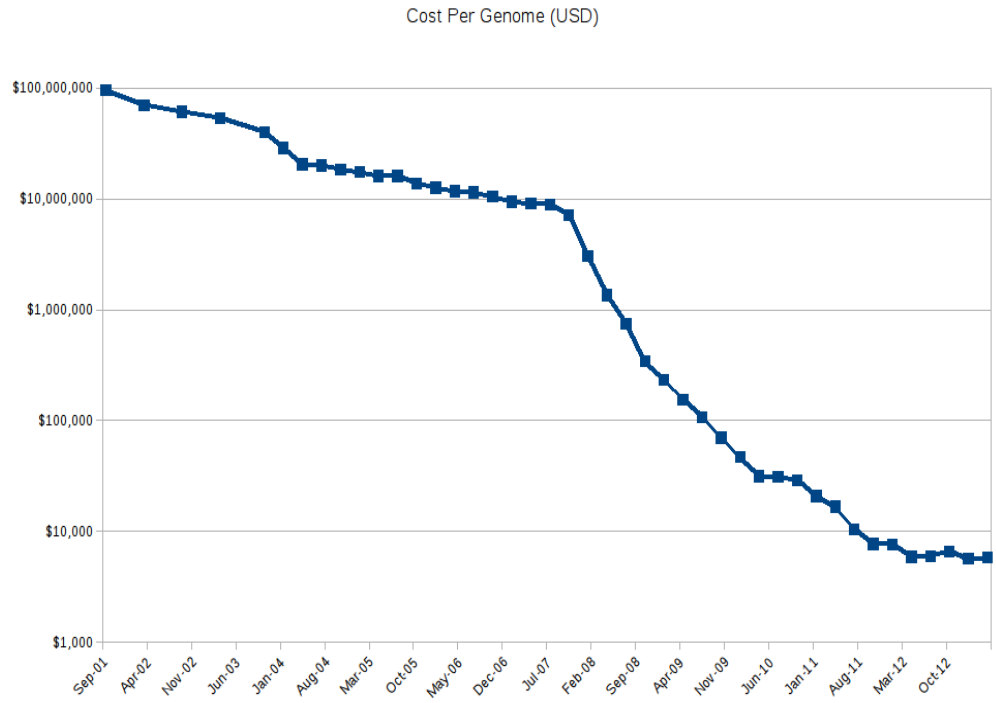

Figure 1.1: Cost Per Genome During the past decade, the cost to sequence a genome has exceeded Moore's Law, dropping from almost $100M in 2001 to under $10K in 2011. Given this pace, the next few years will likely see the achievement of the long-sought $1,000 genome. (Raw data provided by National Human Genome Research Institute and assumes a genome size of 3Gb with 30X coverage using an Illumina platform with 50–100bp reads. Post 2008, a re-sequencing project with an appropriate reference genome is assumed.)

functioning.

1.2 Mechanisms of Regulation

Almost every cell in the human body contains the same basic genetic material, the same blueprint which can be used to generate each of the hundreds of different cell types and thousands of different proteins which allow us to function as we do. The key difference between cells lies in how the flow of genetic information within them is managed, or in other terms, which genes are expressed and which are not. Understanding how gene expression is regulated can provide insights into all areas of cellular physiology, from developmental biology [10, 11] and cell differentiation [12, 13], to regulation of metabolism [14], and mechanisms of disease [15, 16] and ageing [17, 18]. In this section we introduce two layers of regulation – genetic regulation, which we describe in terms of sequence-specific DNA-binding proteins known as transcription factors (TFs), and epigenetic regulation, which

involves various mechanisms including: post-translational modification of histone proteins, chemical 'tagging' (methylation) of CG-dinucleotides, and gene silencing via small non-coding RNAs known as microRNAs (miRNAs).

1.2.1 Genetic Regulation

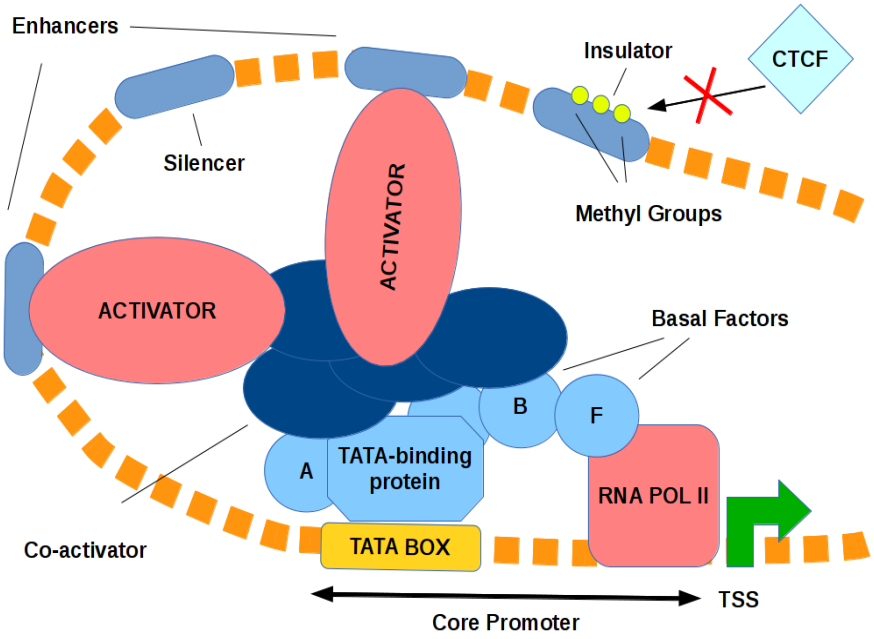

Figure 1.2: Genetic Regulation The initiation of transcription in eukaryotic genomes depends on the presence of RNA polymerase II, a number of basal or general transcription factors, and the binding of sequence-specific transcription factors to enhancer, silencer, or insulator sequences called response elements (REs). Additional proteins functioning as co-activators or co-repressors may also play a role. Shown here is a methylated insulator region (methylation will be further discussed later in this chapter) preventing the binding of the CTCF (CCCTC-binding) transcription factor and thereby allowing transcription to proceed. CTCF plays a diverse role in gene regulation and is largely responsible for chromatin organisation into loop structures by binding to itself as a homodimer [19]. (Source: Based on [20].)

The central dogma of molecular biology as outlined by Crick [21, 22] indicates that, in general, DNA is transcribed to RNA which is then translated into proteins. Since this description assumes that the transcribed gene encodes a protein, more specifically, the DNA will be transcribed to messenger RNA, or mRNA. The process of transcription (Figure 1.2) is mediated both by RNA polymerase and general transcription factors, as well as the previously mentioned sequence-specific

transcription factors[2]. In order for transcription to occur in a eukaryotic cell, RNA polymerase II must, in the presence of these general or basal transcription factors, recognise and bind to a particular sequence of DNA in the core promoter. This is a region located just upstream of a gene's transcription start site (TSS). This complex of general transcription factors and RNA polymerase is known as the preinitiation complex, and although its binding is sufficient for a gene to be transcribed at a low level, expression can be greatly modulated by the effect of sequence-specific transcription factors binding further 5' of the TSS in what is known as the upstream regulatory, or promoter region. Since the binding of these factors can introduce conformational changes in the DNA (DNA looping), distal but functionally relevant binding sites may be located several kilobases away from the TSS.

Transcription factors are a hugely important group of proteins accounting for approximately 8% of the genes in the human genome [23]. They are well-conserved across species [24, 25], and in general, the larger the genome, the more transcription factors it will contain [26]. Transcription factors can both inhibit the transcription rate of a gene (binding to a silencer element and acting as a repressor), or increase the rate of transcription (binding to an enhancer element and functioning as an activator). A single transcription factor can be involved in regulating multiple genes, and individual genes are usually regulated by groups of transcription factors – these higher-order regulatory structures are referred to as *cis*-regulatory modules (CRMs) [27]. Regulatory modules can either help to recruit and stabilise the preinitiation complex [28], or inhibit the binding of RNA polymerase. All transcription factors possess at least one DNA-binding domain (DBD) which allows them to recognise a specific sequence of DNA in the upstream promoter of a gene; common examples are the zinc finger (Zn), basic helix-loop-helix (bHLH), winged helix, HMG-Box, and basic leucine zipper (bZIP) domains (for a more comprehensive list, the reader is directed to [29, 30]). They may also contain additional DBDs, allowing them to recognise more than one binding sequence, as well as trans-activating domains (TADs) which allow them to bind other proteins which function as co-activators or co-repressors – an example of this is the recruitment of either histone acetyltransferases (HATs) or histone deacetylases (HDACs) which serve to alter the association of DNA with histone proteins, making it more or less accessible to the transcription machinery [31] (further discussed in the following section). Transcription factors themselves are of course also subject to regulation, requiring other transcription factors for their expression, or even regulating their own transcription in both positive [32] and negative [33] feedback loops. While some transcription factors such as specificity protein 1 (Sp1) are constitutively active, others require activation before localising to the nucleus for binding; this can be regulated by external stimuli (examples include nuclear factor kappa-light-chain-enhancer of activated B cells, NF-KB, and sterol regulatory element-binding protein 1,SREBP-1), post-translational modification (signal transducer and activator of transcription 1 (STAT1) is activated by phosphorylation), and binding of ligands or other transcription factors to form heterodimers [34].

[2]For the remainder of this work the term transcription factor (TF) will refer to the latter unless otherwise stated.

1.2.2 Epigenetic Regulation

Epigenetics refers to the study of heritable changes in gene expression due to mechanisms which do not alter the basic DNA sequence of an organism (as does, for example, mutation). These changes are maintained throughout cell division [35, 36], and transgenerational inheritance of epigenetic modifications has been observed in a wide variety of organisms [37]. Three main mechanisms of epigenetic regulation are studied – methylation of CG-dinucleotides, post-translational modification of histone proteins, and gene silencing by microRNAs.

Histone Modifications

In order to fit the more than 1.8 metres of DNA found in each human cell into the nucleus it must be tightly packaged. The basic unit of packaging for DNA is the nucleosome, which consists of 147bp of DNA wrapped around a core of histone proteins [38]. This histone core consists of two H2A-H2B dimers and a H3-H4 tetramer. Nucleosomes are further packaged into higher order chromatin structures with the addition of H1 linker proteins. Like all proteins, histones are subject to a variety of post-translational modifications such as acetylation, phosphorylation, ubiquitination, and methylation. Modification of specific residues, particularly in the tails (N-termini) of histones H3 and H4, have been shown to effect changes in the structure of chromatin [9, 39, 40], transforming it from condensed heterochromatin to less tightly packed euchromatin which is more accessible to the transcription machinery. This has lead to the development of a 'Histone Code' [41, 42] which indicates which modifications in which positions constitute active or repressive marks, corresponding to transcriptionally active or transcriptionally silent genes. Examples of active marks include mono- and tri-methylation of histone H3, lysine 4 (H3K4me1, H3K4me3) [43, 44], mono-methylation of histone H4, lysine 20 (H4K20me1) [45], and acetylation of histone H3, lysine 9 (H3K9ac) [44]. Repressive marks include di- and tri-methylation of histone H3, lysine 27 (H3K27me2, H3K27me3) [45] and di-methylation of histone H3, lysine 9 (H3K9me2) [46]. We have previously mentioned that the binding of transcription factors can either directly or indirectly effect a change in the acetylation status of histone tail residues (for example through co-activators such as E1A-binding protein p300/CREB-binding protein, p300/CBP); recent research has demonstrated that lysine acetylation in histones represents a mechanism for targeting the recruitment of bromodomain-containing chromatin remodellers such as those found in the SWI/SNF family [47, 48]. These chromatin remodellers specifically bind acetyl-lysine and further open chromatin by either disassembling and reassembling the nucleosomes, or shifting them along the DNA strand [49] making promoter regions accessible to the preinitiation complex. Similarly, methyl-lysine residues may serve to attract another class of chromatin remodellers containing a chromodomain subunit [50, 51].

CG Methylation

Methylation of DNA involves the addition of a methyl (CH_3) group to cytosine in a CG-dinucleotide (or CpG) context, and is regulated through DNA methyltransferases (DNMTs). Some of these

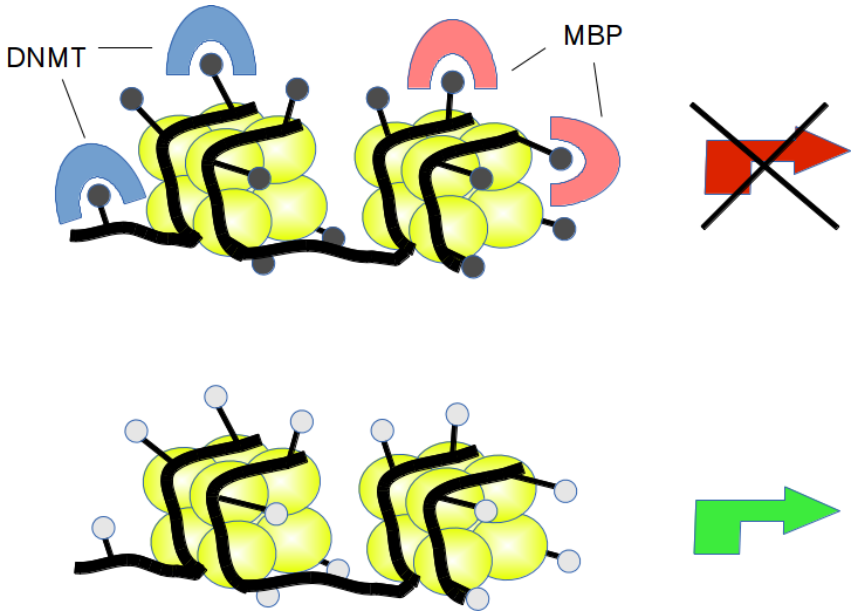

Figure 1.3: Epigenetic Regulation by Methylation Shown here are the histones with methylated (upper panel) and unmethylated (lower panel) cytosines in CG-dinucleotide contexts. The upper panel shows the methylation caused by the DNA methyltransferase enzyme and associated repression of gene expression due to the binding of methyldomain binding proteins (MBPs).

enzymes are responsible for *de novo* methylation (DNMT3a and DNMT3b), while others serve to maintain methylation on newly created strands during DNA replication (DNMT1). DNA methylation plays an important role in many vital cellular processes such as genomic imprinting [52] and X-chromosome inactivation (XCI) [53]. Because 5-methylcytosine can spontaneously deaminate to thymine, an occurrence that can result in a transition mutation [54], CG-dinucleotides tend to be underrepresented in organisms which methylate their DNA [55]. Although methylation can be widespread (mammals have been shown to methylate up to 90% of the CGs in their genomes [56]), an important set of CG-rich regions exist which demonstrate protection from methylation. These regions, termed 'CpG islands', are located in approximately 40% of mammalian promoters [57] and are defined as regions of greater than 500bp in length with a GC content greater than 55% and an observed to expected CpG ratio of 0.65 [58]. Transcriptionally active genes will usually have unmethylated CGs in their promoters, while methylated CGs in the promoter region result in transcriptional repression. Repression may either be due to the fact that transcription factors cannot bind to the methylated cytosines, or that their binding is blocked as a consequence of methyl-CpG-binding-domain proteins (MBDs) such as methyl-CpG-binding domain protein 2 (MBD2) or methyl-CpG-binding protein 2 (MECP2) already being bound (Figure 1.3). MBDs are also known to recruit co-repressor complexes including histone deacetylases, presenting a further mechanism for gene silencing [59]. Methylation of CGs can help to protect a cell from the effect of poten-

tially harmful transcripts such as transposable elements and oncogenes; evidence of this is provided by studies of methylation patterns in cancer phenotypes which have shown both hypomethylation of normally-methylated oncogene promoters such as c-Myc (v-myc avian myelocytomatosis viral oncogene homolog), as well as hypermethylation in normally-unmethylated tumor suppressor gene promoters, such as tumor protein 53 (p53) [60, 61, 62]. Because these genes encode transcription factors, their aberrant expression can cause dramatic changes to the expression of all downstream genes, resulting in severe disruption of the normal phenotype.

MicroRNAs

MicroRNAs (miRNAs) are small (21–25nt) non-coding RNAs which negatively regulate the expression of their target mRNAs by binding with perfect or near-perfect complementarity to their 3' UTRs, resulting in either cleavage of the mRNA [63], inhibition of translation [64], or targeting of the mRNA for degradation through de-adenlyation [65].

They were first discovered in *C. elegans* by Lee et al. in 1993 [66] and since then have been shown to have an important role in the regulation of many different functions including development [67, 68], cell differentiation [69, 70], apoptosis [71], and metabolism [72]. They have also been implicated in the pathology of several diseases such as cancer [73, 74, 75], heart disease [76, 77], and neurological disorders [78, 79]. This evidence, coupled with the fact that miRNAs are evolutionarily well-conserved [80] and that there are approximately 1,000 miRNAs in the human genome[3] targeting up to 60% of mRNAs [81], identifies them as an important class of biomarkers and potential therapeutic targets [82].

MiRNAs can be found in intergenic regions, being independently transcribed, or exist in the introns (and even exons) of both protein-coding and non-protein coding genes where they are transcribed along with the host gene [83, 84]. The first step in their synthesis (shown in Figure 1.4) is the transcription of a primary miRNA (pri-miRNA), which may contain multiple precursor-miRNAs (pre-miRNAs) in the form of stem-loop structures. The pri-miRNA is then processed by the dsRNA-specific ribonuclease Drosha [85] which releases the pre-miRNA hairpins for export to the cytoplasm. Once in the cytoplasm, the pre-miRNA is cleaved by the Dicer enzyme, resulting in an miRNA:miRNA* duplex [86] which then separates into the mature miRNA (or guide strand) and the second (or passenger) strand. The mature miRNA is incorporated into the RNA-induced silencing complex (RISC complex) for interaction with its mRNA target, while the second strand is usually degraded (although this is not always the case [87]). While either strand from the duplex may become the active miRNA, it is thought that the less stable strand may preferentially associate with the RISC complex [88].

1.2.3 Regulatory Complexity

While we have provided only a brief overview of some genetic and epigenetic mechanisms of regulation, it should be clear that the complexity involved in these systems is quite large. Much of

[3]Mirbase – http://www.mirbase.org/

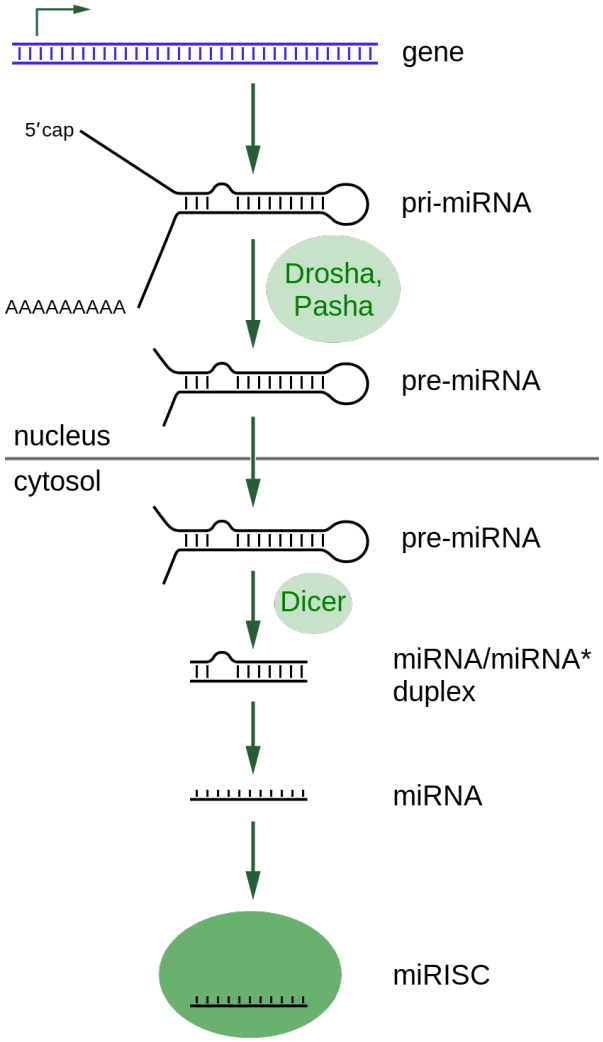

Figure 1.4: MicroRNA Processing Shown here are the various steps in the processing of microRNAs from initial transcription through cleavage by Drosha and Dicer and finally, incorporation into the RISC complex. (Source: Wikimedia Commons, public domain.)

this complexity stems from the fact that all of these mechanisms are interconnected – transcription factors can regulate the expression of methyltransferases, miRNAs can target transcription factors, miRNA promoters may be methylated, transcription factor binding can recruit histone modifying proteins which may in turn recruit chromatin remodellers, and so on. This connectivity helps to create the complex regulatory networks necessary to produce the range of cell types and functions commonly found in multicellular organisms.

1.3 Thesis Outline

This work focuses on the use of machine learning algorithms, automated pipelines, and high-performance computing for the analysis of genome-scale regulatory data. In Chapter Two, we introduce the key concepts of NGS and describe a platform developed to facilitate the automated processing of data at the recently-created Center for Epigenomics at the Albert Einstein College of Medicine. This system, the Wiki-based Automated Sequence Processor, or WASP, has been instrumental in allowing researchers at Einstein to leverage the power of new sequencing technologies and provides both infrastructure and algorithms to manage the primary analysis of the most common sequencing-based assays. By automating the storage, analysis, and return of experimental data generated from MPS-based experiments, researchers are free to concentrate on the biological question or hypothesis rather than spending time learning to cope with the massive amounts of raw sequencing data.

Having covered primary analysis in Chapter Two, in Chapter Three, we describe the application of two machine learning techniques to the secondary analysis of data related to the genome-wide study of transcription factors. We introduce some of the various ways in which DNA-binding sites can be represented and then discuss the problem of *de novo* motif finding. We outline the previously published SOMBRERO algorithm [89], which uses the self-organising map (SOM) neural network (NN) [90] to identify enriched sequence motifs and describe some of its key limitations. We then provide solutions to some of these limitations based on our modifications to the original algorithm and present the resulting implementation, ChIPSOM, demonstrating its improved scalability and application to ChIP-chip data. We further discuss the issue of redundant motif predictions and show how this limitation may be overcome by a novel secondary clustering approach using a genetic algorithm (GA) [91, 92]. We demonstrate the effectiveness of our novel algorithm, GMACS, on the more general problem of the automated construction of familial binding profiles (FBPs).

In Chapter Four, we describe a study to explore the role of the T-Box 1 (TBX1) transcription factor in postnatal neuronal development and the effect of Tbx1 haploinsufficiency on murine social behaviour and vocal characteristics. We provide details on the primary and secondary analysis of the ChIP-seq data using the earlier described WASP, ChIPSOM, and GMACS software. We combine this genetic analysis with a phenotypic analysis using an entropy-based tool we have developed called Mumbles, as well as some further statistical machine learning techniques, to determine inherent structure in strings of mouse vocal calls.

Finally, in Chapter Five, we present a summary of our work and briefly discuss some general conclusions and potential future directions.

CHAPTER 2

Automated Primary Analysis of Next Generation Sequencing Data

The work in this chapter was carried out in collaboration with members of the Computational Genomics and Epigenomics Groups at the Albert Einstein College of Medicine. The candidate was involved in the overall design of the WASP system and had specific responsibility for the frontend development including MediaWiki customisations and extensions, as well as programming of the AJAX/PHP sample submission forms facilitating the capture of sample data and metadata to the MySQL database. The candidate also developed code for the system backend, including initial versions of the 'watcher' script responsible for monitoring of run-related folders and invocation of assay-specific processing pipelines.

2.1 ChIP-seq

In this section we introduce the key concepts of sequencing-based assays. Although there are many sequencing platforms currently available from vendors such as Roche/454 Life Sciences, Illumina/Solexa, Life Technologies, Oxford Nanopore, and Pacific Bioscience, each with their own advantages and disadvantages; here, we will focus on the Illumina platform (and in particular the Genome Analyzer IIx, or GAIIx system) since, at the time of writing it is: 1) currently the most prevalent platform[1], and 2) the workhorse machine of the Einstein Center for Epigenomics, which will be discussed later in this chapter. We use ChIP-seq as an example NGS assay as it will be further discussed in terms of secondary analysis in Chapter Three, and its application in basic neuroscience research is described in Chapter Four.

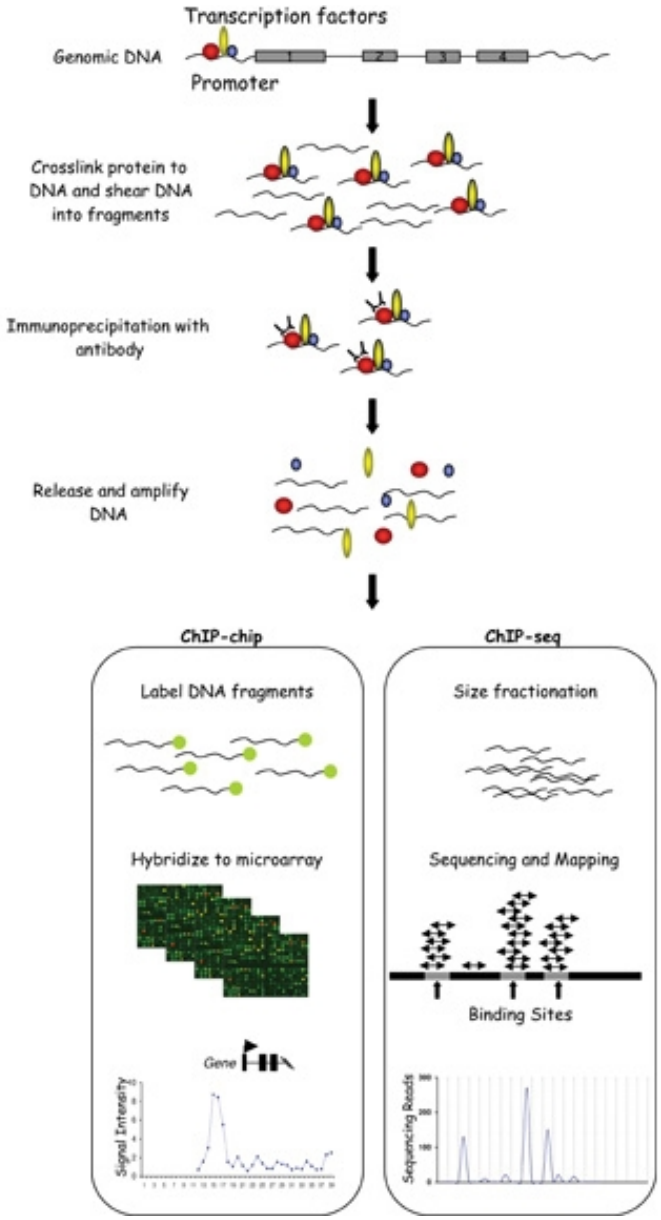

Figure 2.1: High-Throughput ChIP Workflows The DNA is first treated with formaldehyde in order to cross-link transiently-bound proteins and is then sonicated to produce low molecular weight fragments. Antibodies specific to the protein of interest (POI) are added to enrich for protein-DNA complexes and unbound DNA is washed off. The cross-links are then reversed by heating, and in the case of ChIP-chip, the eluted DNA is purified, amplified, and labelled using fluorescent tags and hybridised to a microarray. The signal from the fluorescent tags identifies genomic regions corresponding to binding sites. In the case of ChIP-seq, libraries for the NGS platform of choice are created and the resulting sequence reads are mapped to a reference genome resulting in location-specific binding peaks. (Source: Wikimedia Commons. Adapted from [93] and licenced under CC-BY-3.0)

Introduction

ChIP-seq [94] combines chromatin immunoprecipitation with Massively-Parallel Sequencing (MPS) in order to produce much higher resolution binding data, with less noise and at greater coverage than is currently possible with microarray-based approaches [95]. Sequencing offers the advantage of avoiding all of the complications inherent in array hybridisation such as probes with different optimal binding temperatures, non-specific hybridisations, DNA secondary structure interfering with hybridisation, and so on. Massively-parallel sequencing also offers a more cost-effective way to generate genome-wide biding data than, for example, tiling arrays [96].

The ChIP-seq protocol begins in a manner similar to that of ChIP-chip (Figure 2.1), but instead of washing labelled DNA over a microarray, both the immunoprecipitate (IP) and input (unprecipitated DNA used as a control) have adapters ligated which facilitate Illumina's bridge-amplification and sequencing-by-synthesis approach (described in Figure 2.2). Illumina offers both single- and paired-end library preparation and sequencing. In the latter, the DNA molecule is sequenced from both ends allowing both higher mapping accuracy (including the ability to assemble reads which are mapped across repetitive regions) and the identification of structural rearrangements such as insertions, deletions, and inversions [97]. Samples may also be multiplexed using index sequences, allowing 12 samples to be run on one lane, or 96 samples per flowcell – this makes the GAIIx particularly cost-effective when performing targeted sequencing of specific genomic regions, such as in the case of exome sequencing.

Although we present the platform in terms of the ChIP-seq assay, the GAIIx system also has applications in many other areas such as RNA-seq (including sequencing of miRNAs and other small RNAs), and with read lengths of 150bp now available, is increasingly being used for *de novo* sequencing and whole-genome and targeted resequencing. It is also possible to perform genome-wide epigenetic profiling using any of the many available methylation-based assays such as MeDIP-seq [98], Bisulfite Sequencing [99], or HELP-tagging [100].

Data Analysis

ChIP-seq primary data analysis consists of three main stages – basecalling, where the images captured during each cycle are analyzed to determine which nucleotides were incorporated, read mapping, where the sequences generated are mapped to a reference genome, and peak calling, where the clustered reads, or peaks, are identified as significant (corresponding to a likely binding site) or not. We will focus on the last two stages since the basecalling is now carried out in real-time on the instrument itself as the flow cycles progress. This real-time processing and discarding of raw image files is one of the ways in which both manufacturers and institutions have tried to address the massive storage issues facing sequencing facilities.

Although Illumina provides its own short read mapping algorithm, Eland, as part of its internal pipeline, there are a variety of other aligners available which usually seek to optimise either speed,

[1] World Map of High-throughput Sequencers – http://www.pathogenomics.bham.ac.uk/hts/

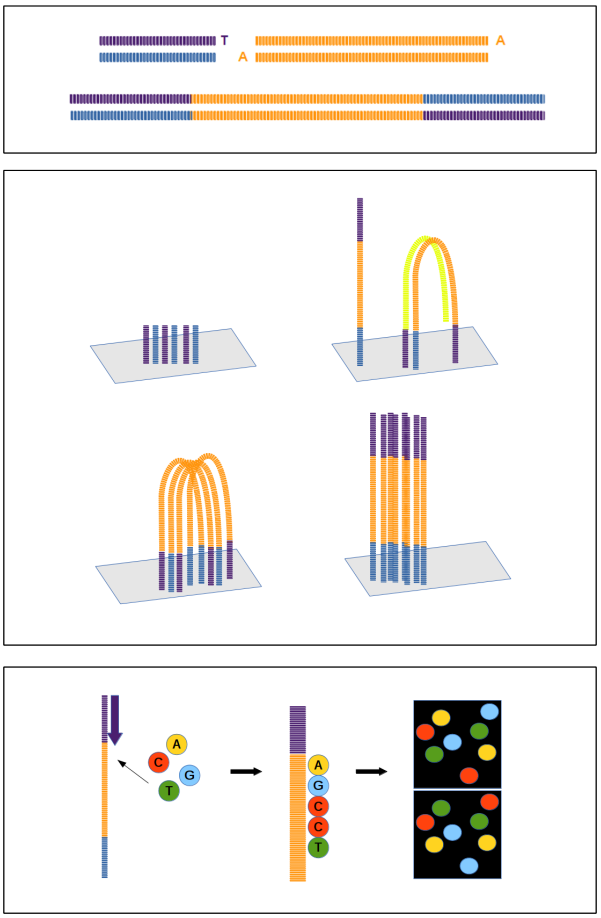

Figure 2.2: Sequencing Process An overview of the Illumina GAIIx sequencing process from library preparation (upper panel), through cluster generation (middle panel), and nucleotide incorporation and basecalling (lower panel). Once the POI has been precipitated, the ends of the eluted DNA are blunted and adenolated so that Illumina adapters (shown in purple and blue) containing a sticky T overhang can be ligated. Size-selection via gel extraction is then used to isolate fragments in the desired range (usually 150-250bp). These fragments are run on a flow cell (shown in grey) which is coated with a lawn of primers. Once bound, millions of copies, or clusters, are created through repeated extension and denaturation. Finally, in each sequencing cycle, fluorescently labelled nucleotides are flowed across the cell and when one is incorporated, the resulting fluorescence is captured by a high-quality imaging system. As the sequencing, or flow cycles are repeated, each fragment sequence is extended by one base, resulting in a sequence of images in which the newest incorporated base in each cluster is visible. Once the images have been processed (converting fluorescence to called bases), the resulting reads are mapped back to a reference genome to identify the locations where the protein of interest was bound in the original sample. (Based on an image from http://illumina.com)

accuracy, or memory footprint when mapping reads to a reference genome; below, we provide some examples.

Maq [101] is an algorithm specifically designed for Illumina/ABI-SOLiD reads that uses a concept of mapping quality based on the Phred quality score [102]. The mapping quality score provides a measure of confidence that a read actually comes from the position to which it is aligned in the target. Maq first indexes all of the reads, then scans through the reference genome several times, storing the best hits, allowing for a small number of mismatches. It has the ability to perform gapped alignments for paired-end reads, can handle read lengths of up to 63bp and, depending on coverage, can call single nucleotide polymorphisms (SNPs). It is however, much slower than Eland, sacrificing speed for quality. It has also been argued that since Maq does not support gapped alignment for single-end reads, it is unsuitable for alignment of longer reads where indels may occur frequently [103].

SHRiMP (the SHort Read Mapping Package) [104] is a tool designed to allow mapping of reads to highly polymorphic genomes. The algorithm first calculates a hash of all spaced k-mers (or seeds) in the reads and then scans the genome using a sliding window approach. Reads having multiple seed matches within the genomic region bounded by the sliding window are then fully aligned using a fast vectorised Smith-Waterman alignment [105]. When used for analyzing reads generated for the resequencing of a *Ciona savignyi* genome, SHRiMP was shown to identify 5-fold more SNPs than the ABI-SOLiD's default mapper, while also capturing 70,000 variants.

Bowtie [106] is a popular aligner, which, unlike Maq and SHRiMP, indexes the reference genome rather than the reads using the Burrows-Wheeler transform [107] as an indexing strategy. This allows Bowtie to perform exact matches while using a backtracking algorithm to account for mismatches. Bowtie is one of the fastest alignment algorithms, having been shown to align 35bp reads at a rate of more than 25 million reads per CPU-hour, more than 35 times faster than Maq under similar conditions. It also has a small memory footprint, with the indexed human genome taking up approximately 1.3GB of system memory during alignment.

The fast, lightweight BWA also uses the Burrows-Wheeler transform to perform alignment of both short [103] and long [108] reads. The short read algorithm performs gapped global alignments on reads of up to 200bp, while the second algorithm uses a heuristic Smith-Waterman-like alignment to align longer reads which may contain more sequencing errors. BWA also supports paired-end reads and provides mapping quality scores.

Mapping millions of short reads to a reference genome is a computationally expensive task; as the ability to generate a greater number of longer reads per sample increases and sequencing of paired-end reads becomes standard, high-performance computing resources become increasingly critical for data analysis. Most read mapping programs therefore either make use of pthreads (or the higher level abstraction OpenMP) to run on multi-core symmetric multiprocessor architectures (SMPs) or are explicitly written using MPI for execution on distributed memory compute clusters. Recent years have also seen the first use of cloud computing for computationally intensive bioinformatics tasks. Cloudburst [109] is a parallel read mapping algorithm based on RMAP [110]. It uses the open-source Hadoop implementation of Google's MapReduce [111] to parallelise execution using

multiple compute nodes and shows near-linear speedup as the number of processors is increased. The potential of the cloud to provide scalable storage for massive amounts of data as well as on-demand, configurable compute clusters is an attractive option for researchers faced with ever-growing high-performance computing (HPC) resource costs; future algorithms will likely be developed with this fact in mind and it may well be that much of the bioinformatics analysis carried out in the coming years will be cloud based.

Once the sequenced reads have been mapped to the reference genome, peaks must be identified and analyzed to determine their statistical significance. As in the case of alignment tools, there are a plethora of algorithms available for this task, we will mention only some of the more popular ones here; for a more comprehensive discussion and comparison of these and other peak calling algorithms, the reader is directed to [112] and [113].

PeakFinder was first described in [94] where it was used to identify neural-restrictive silencer factor/RE1-silencing transcription factor (NRSF/REST) binding sites in Jurkat cells based on the concept of tag clustering. It determined positive binding loci as regions where 13 or more independent sequence reads occur within a distance of 100bp, having at least five partly overlapping reads and demonstrating at least a 5-fold enrichment in the IP when compared with the input.

A two-pass approach is adopted by PeakSeq [114] in determining potential binding sites. The first pass identifies candidate loci by comparison of the IP reads to a simple null background model, while the second pass determines IP enrichment relative to the control. The reads are first extended by the average fragment length and a count of the number of overlapping DNA fragments at each nucleotide position is calculated. The analysis then proceeds on a chromosome by chromosome basis, using 1Mb segments to capture genomic variability and correcting for mappability based on their previously determined mappability map. Candidate binding regions are then identified based on a threshold designed to satisfy a specific false discovery rate (FDR). Before comparing the selected potential binding regions to the input sample to determine enrichment, a normalisation is carried out on the input DNA. A linear regression of the tag count from the input is performed against the IP in 10kb windows along each chromosome and the slope is then used to scale tag counts in the input. Statistical significance is calculated using the binomial distribution and correction for multiple hypothesis testing is carried out by applying a Benjamini-Hochberg correction to the determined p-values.

SISSRs [115] is an algorithm which makes use of the fact that since reads represent the ends of sequenced DNA fragments, forward and reverse strand reads will cluster in overlapping peaks on either side of a binding site. SISSRs scans the genome using a sliding window of 20bp width and subtracts the number of antisense reads within the window from the number of sense reads, identifying binding sites as the transition points of this count from positive to negative. If a control sample is available, an FDR is determined as the peak ratio in the IP versus the control. If no control is present, then a Poisson background model is used.

Another algorithm which takes advantage of this bimodal pattern is MACS [116], which empirically models the shifting size in order to improve the spatial resolution of predicted binding sites. It employs a window-based scanning approach to detect peaks with an m-fold enrichment and then

uses 1,000 of these detected peaks to determine the optimal shift size. To capture small tag biases between different regions in the genome, a dynamic Poisson distribution is used to model a local background. Peaks with p-values below a user-defined threshold are identified as potential binding sites. If negative control data are available an FDR is estimated.

CisGenome [117] is a powerful tool, offering both command-line and GUI interfaces and providing the ability to handle both ChIP-chip and ChIP-seq data as well as providing downstream motif analysis. For ChIP-seq data, it scans the genome with a sliding window to identify regions with enriched read counts. FDRs are estimated assuming that the background read occurrence follows a negative binomial distribution. The authors demonstrate that this type of distribution can provide a better fit to the real data than the global Poisson distribution. The FDR is determined by calculating the ratio between the number of peaks expected by the null model at a particular cut-off level and the observed number of peaks detected at the same level.

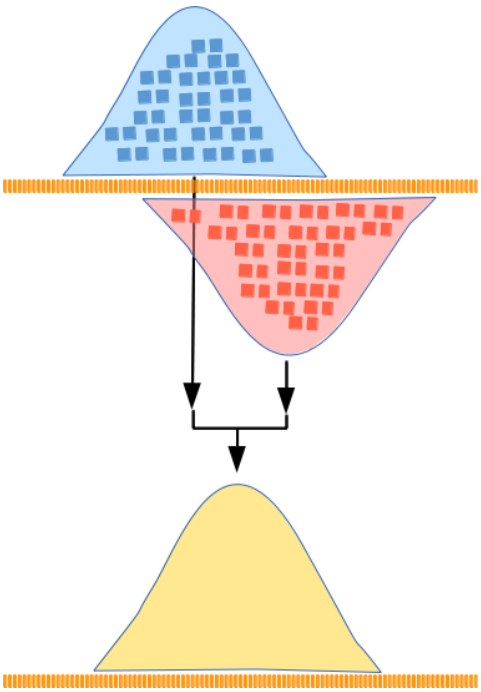

Figure 2.3: Peak Shifting Several approaches make use of the bimodal pattern resulting from reads mapping to forward and reverse strands on either side of the true binding site. The QuEST algorithm calculates a combined density profile (CDP, shown here in yellow) based on a peak shift calculated from applying separate Gaussian kernels to reads on the forward (blue tags) and reverse (red tags) strand.

A slightly different approach is employed by the QuEST [118] algorithm which applies a Gaussian kernel separately to reads from both strands, then calculates the peak shift to form a combined density profile (CDP), the local maximum of which corresponds to the predicted binding site (Figure 2.3). To estimate an FDR, QuEST separates the negative control data into two sets, one of which is used as a pseudo-ChIP sample in which peaks are predicted and the other of which is used as

a background for this sample, the FDR is then calculated as the ratio of peaks predicted in the pseudo-ChIP analysis to the number of peaks identified in the real ChIP experiment. A potential disadvantage of QuEST is that it does not support analyses that do not include control samples.

The Hpeak [119] algorithm provides yet another approach, identifying peaks using a two-state (binding sites and background) hidden Markov model (HMM). The emission probabilities are described by two different Poisson distributions and the significance of enrichment of the peaks is adjusted using the Bonferroni correction for multiple testing.

2.2 The Challenge for Core Facilities

The previous section provided an overview of the algorithms and compute and storage demands for an example sequencing-based assay. With improvements to sequencing chemistry, imaging systems, and associated basecalling algorithms continuously emerging however, NGS platforms are consistently producing increasing amounts of raw data for the same basic sequencing cost (Figure 2.4). A prime example of this is Illumina's recent platform – the HiSeq 2000, which has been promoted as the first commercially available sequencer to enable researchers to obtain ∼30x coverage of two human genomes in a single run for under $10,000 per sample[2].

	Read Length	Run Time	Output
GA II x	1 x 35 bp	~2 days	10 – 12 Gb
	2 x 50 bp	~5 days	25 – 30 Gb
	2 x 75 bp	~7 days	37.5 – 18 Gb
	2 x 100 bp	**~9.5 days**	**54 – 60 Gb**
	2 x 150 bp	~14 days	85 – 95 Gb
	Clusters passing filter – 320m / 640m paired-end		
	Throughput - Up to 6.5 Gb per day for a 2 x 100 bp run		
HiSeq1000	1 x 35 bp	~1.5 days	13 – 17.5 Gb
	2 x 50 bp	~4 days	37.5 – 50 Gb
	2 x 100 bp	**~8 days**	**75 – 100 Gb**
	Clusters passing filter – 500m / 1billion paired-end		
	Throughput - Up to 12.5 Gb per day for a 2 x 100 bp run		
HiSeq2000	1 x 35 bp	~1.5 days	26 – 35 Gb
	2 x 50 bp	~4 days	75 – 100 Gb
	2 x 100 bp	**~8 days**	**150 – 200 Gb**
	Clusters passing filter – 1billion / 2billion paired-end		
	Throughput - Up to 25 Gb per day for a 2 x 100 bp run		

Figure 2.4: Illumina Sequencing Platforms Statistics on three different Illumina platforms give an indication of the processing time and amount of raw reads generated during paired-end 100bp runs. An almost four-fold increase in throughput is evident when moving from the GAIIx to the HiSeq2000 platform.

As these sequencing costs decrease and more and more investigators adopt these assays as standard discovery and diagnostic tools in basic molecular genetics and translational science, sequencing has ceased to be the remit of a few select dedicated centres and has instead become a service that is being

[2]http://www.illumina.com

provided at most research institutions by individual labs or core facilities. This shift necessitates streamlined bioinformatic support for the analysis, presentation, and integration of the massive amounts of sequence data generated by these platforms. In the remainder of this chapter, we present the WASP system which was designed to address these challenges.

2.3 The WASP System

In[3] 2009, the Epigenomics Shared Facility (ESF) at Einstein was moving to full production mode, offering sequencing assays to both internal and external researchers. Unfortunately the ability of the new core to process and analyze massive sequencing datasets was lagging severely behind their ability to generate them. At that time, raw sequence data were returned to investigators either as links to an FTP site or on a physical medium, leaving them to rely on their own bioinformatics experience to fully explore their results, or to enlist the help of either a collaborator, or a dedicated fee-for-service bioinformatics core. As indicated in [121], the bottleneck in genomic discovery at this stage becomes not the sequencing itself, but rather the processing and analysis of the generated data. In order to enable high-throughput automated analysis of the data generated by the core, a system was needed that provided sample submission, core facility laboratory management, primary data analysis and return of result to investigators in a user-friendly manner, as well as providing billing and administrative oversight. Thus the development of the Wiki-based Automated Sequence Processor, or WASP, system was begun. This system, which makes use of a variety of software technologies as well as a dedicated HPC cluster has been developed in collaboration with members of the Computational Genomics and Epigenomics Groups at the Albert Einstein College of Medicine, and has been deployed since late 2009 as part of the Einstein Center for Epigenomics, where it currently serves the needs of a large research community employing MPS-based assays for both genetic and epigenetic experiments. The system processes 4–6 terabytes of raw data per day and has processed well in excess of 1 petabyte since its initial deployment.

2.3.1 Introduction and Architectural Overview

We describe the WASP system in terms of three main components: 1) User-side interface, which includes user registration, sample submission and return of processed data, 2) Laboratory information management system (LIMS)/administrative interface which allows core facility personnel to set up and track sequencing runs as well as perform billing and administrative functions, and 3) Backend processing, which details the databases, scripts, and pipelines created to handle each of the distinct assays which the system can process. An overview of the basic architecture of the WASP system is shown in Figure 2.5.

[3]Parts of this section have previously been published in [120] and appear with permission. Unless otherwise specified, figures are screenshots from the WASP system – http://www.wasp.einstein.yu.edu.

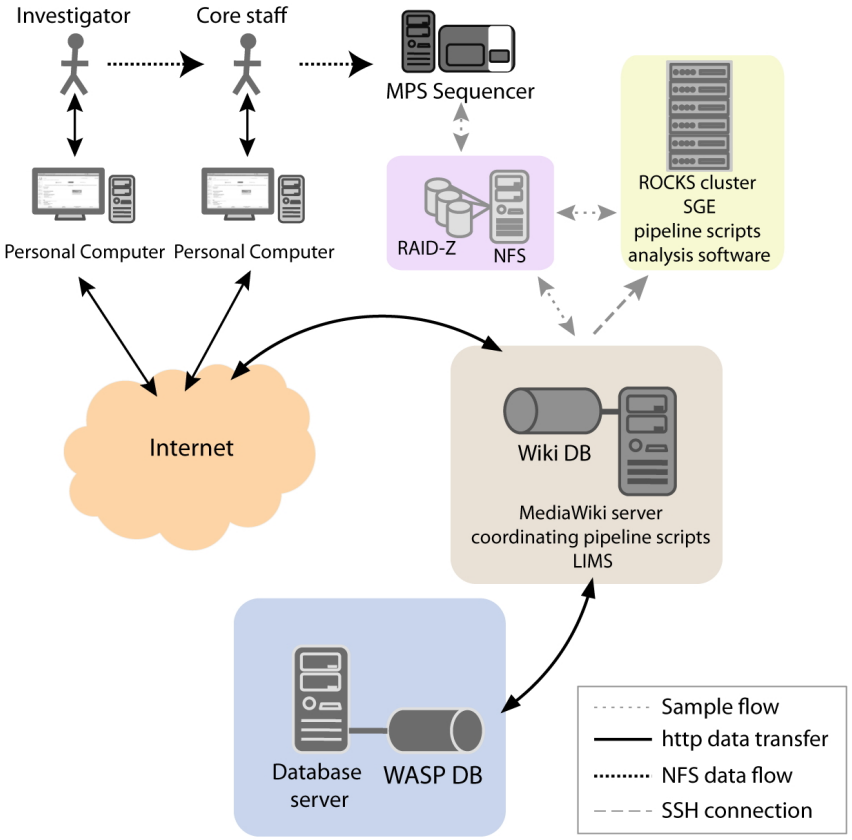

Figure 2.5: WASP Architecture The WASP system comprises three conceptual layers: 1) The presentation layer is primarily provided by an instance of MediaWiki, the popular PHP-based open source wiki package, but also includes a custom-built AJAX sample submission system, 2) The data layer includes MySQL databases which handle data persistence for both the MediaWiki instance and the custom-designed core facility LIMS system as well as an network file system (NFS) disk array for raw and processed sequencing data, 3) The logical layer is implemented on the main WASP server and cordinates processing tasks using a variety of programming and scripting tools which will be discussed in a later section. The main server also functions as a submit node to a local HPC resource allowing parallelised data processing and analysis. This resource entails a 1,360-core Rocks cluster running Sun Grid Engine (SGE), a 72 TB write-optimised RAID-Z disk array for raw data storage and a 42 TB read-optimised RAID-Z NFS disk array for data processed by the various WASP pipelines. All hosts, scripts, and databases are regularly backed up and completed sequencing runs are permanently archived to tape after 3 months. Once an investigator delivers the sample to the core facility, they interact with the system via the Wiki interface to receive updates on the various stages of sample processing as well as all sequencing results. The core facility staff set up run information and provide metadata through the dedicated ESF LIMS component. The main server is then responsible for submitting analysis jobs to the HPC cluster and updating the Wiki and MySQL database as necessary. (Source: [120])

2.3.2 User Interface

A wiki was chosen as the front end for the WASP system for two main reasons. Firstly, it comprises an advanced content management system which includes features such as versioning and data integration (i.e. the ability to easily combine text, images and hyperlinks to external resources in a wikipage). This combination of features not only provides flexibility in terms of data presentation, but also enables a user to maintain a timestamped electronic multimedia lab notebook, a useful concept in terms of validating scientific discoveries relating to intellectual property (IP) claims. Secondly, the wiki interface provides users with an immediately familiar environment which empowers them to create, edit, and share information in an intuitive way using simple tools. This has proved to be invaluable for collaborative wikipages on topics such as protocol design and optimisation.

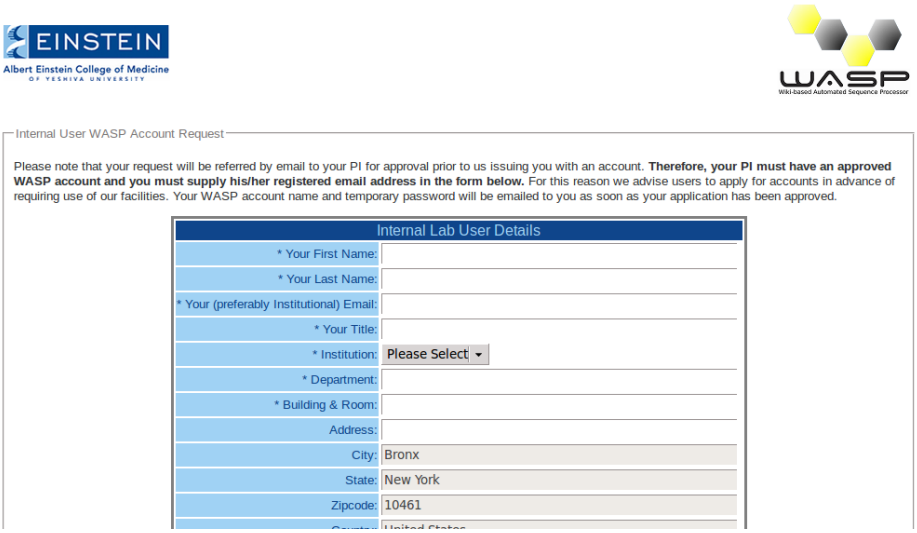

Figure 2.6: WASP System Registration Users can register as either a new PI, invoking the creation of a new lab group, or as members of an existing lab to which they will be added. The WASP system also allows differentiation between internal and external job submissions for administrative, billing, and reporting purposes.

MediaWiki[4] is an open source wiki package written in PHP originally designed for use on the popular Wikipedia[5] project. We have added access control extensions to the base installation which allow us, in a flexible manner, to control access to individual pages on a user or group basis. This ability to restrict viewing and editing of pages to an individual investigator or lab group while also providing access across multiple investigators or groups for collaborative projects is essential for achieving our design goal of 'secure collaboration'.

To enable this level of access control, all users of the system are required to register using our automated registration process. During registration (Figure 2.6), users indicate to which lab they are affiliated and will only be authorised when the Principal Investigator of that lab clicks a link in

[4]http://www.mediawiki.org/wiki/MediaWiki
[5]http://www.wikipedia.org/

an email which is automatically sent to them in response to the user's request to be added to the wiki lab group. Once authorised, login details are emailed to the user and a personal wikipage is created for them. The user's wikipage contains links to data and results for all of the projects for which they have submitted samples as well as a 'Submit New Sample' button which moves them to our custom sample submission system.

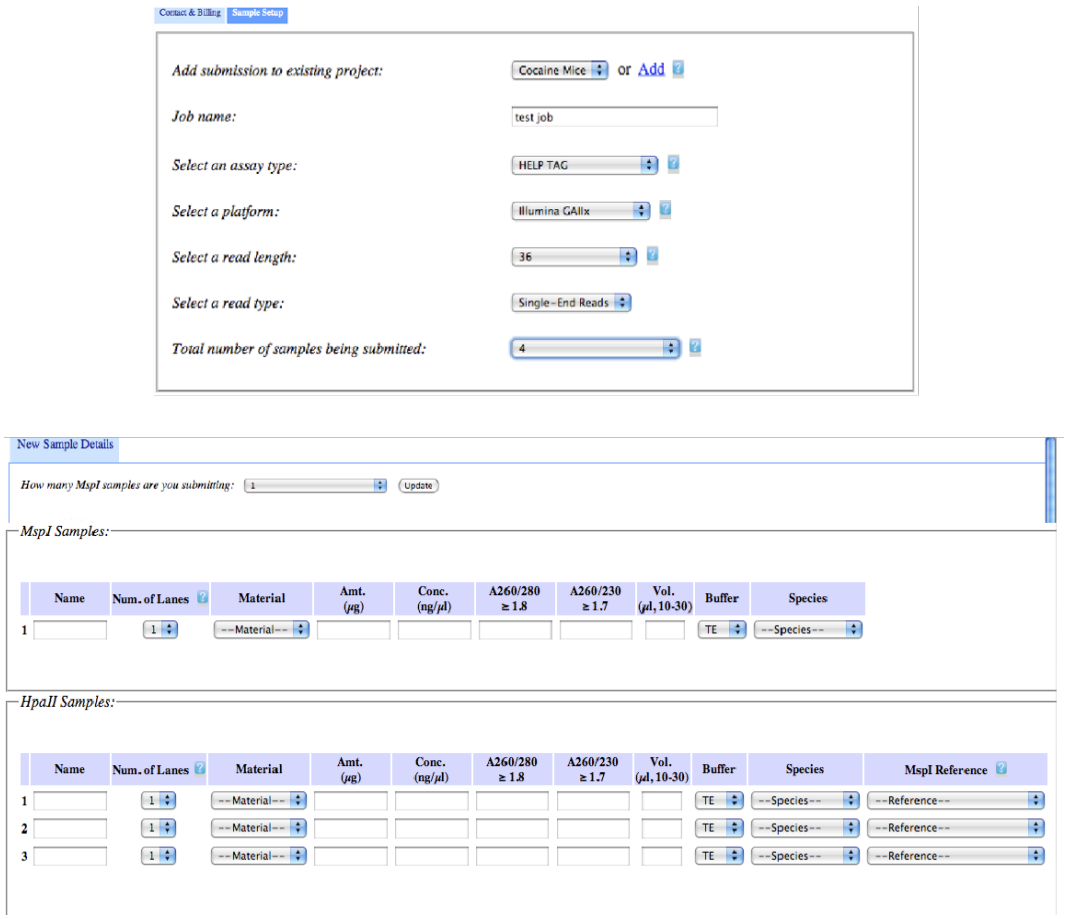

Figure 2.7: AJAX Submission Forms Dynamically generating and pre-populating form elements ensures fast, easy sample submission and greatly reduces costly user input errors. In this HELP-tagging example: 1) The project selector allows a user to add this job to an existing project or use the provided link to create a new project, 2) The assay selected creates a new form section populated with appropriate sequencing platform choices for this assay, 3) The selected platform results in a new form section being generated with read-length and paired-end status selectors associated with this platform, 4) Combining information from the total number of samples as indicated in the sample setup tab and number of MspI references selected in the sample details tab, the appropriate number of form fields are generated for providing details on HpaII samples (including drop-down options for linking samples.)

Given that the core facility can perform a range of different assays on multiple platforms (with

each assay requiring distinct metadata), we were eager to ensure our sample submission system would not be overly complicated or confusing for users. We also aimed to keep the time required to submit sample information to a minimum. We therefore developed an AJAX (asynchronous Javascript plus XML)-based system which dynamically builds modular submission forms in response to user input. Depending on details such as whether or not this is a new project, if this project uses a new or existing grant, platform type, assay type, number of samples and so on, appropriate form elements will be automatically generated to accept user input (Figure 2.7). Where possible, we also pre-populate form elements with data retrieved from the database based on previous user submissions so that commonly used reagents, antibodies and primers can be selected rather than re-inputed. Most selectors also include a tooltip (indicated as a blue question mark icon to the right of the drop-down menu), which when hovered over, will provide details and/or advice on selection choices. This ensures that users can quickly and easily provide details on their specific submission, while allowing us to capture all of the data and metadata necessary to correctly process their samples. In order to ensure appropriate QC for sequencing runs, users are required as part of the submission process to upload either gel images or Bioanalyzer output, as well as PCR primers which core staff can use to check for enrichment.

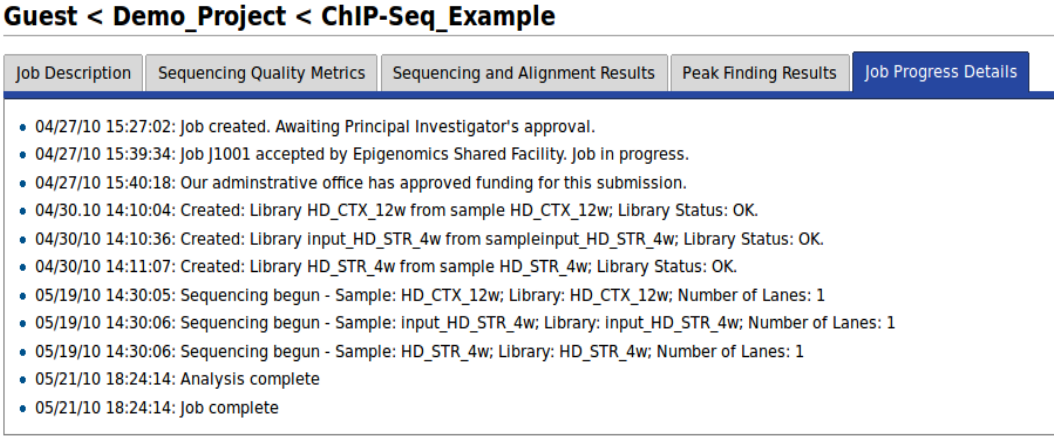

Figure 2.8: Job Progress Details Tab This wikipage is automatically updated in response to different stages of the job workflow being either initiated or completed. Allowing users to track job progress through their wikipages greatly reduces the number of inquiries from WASP users and therefore reduces the associated burden on core staff.

Once all appropriate sample metadata has been entered, the PI of the lab is automatically sent an email with links to either approve or withdraw the submission. If approval is given, a confirmation email, including an automatically generated quote in PDF format is sent to the job submitter, the PI, and to a core administrator who then verifies that the account indicated by the grant number provided as part of the submission contains appropriate funds. Once this has been verified, and the samples have been submitted to the core, generation of the sequencing libraries can proceed. After samples are submitted, the user can see, in real-time, the progress of their submission using the 'Job

Progress Details' tab in the job wikipage (Figure 2.8). When events such as approval of funding, creation of libraries, completion of sequencing and beginning and end of analysis occur, the WASP system automatically updates this progress page with a timestamped record of the event.

When the analysis of submitted samples has been completed, another automated email is sent to the user with a link to the wikipage under their account which contains the job results. Depending on the assay performed, different tabs will be included as part of the results page. The first of these tabs is the Job Description tab (Figure 2.9), which provides a summary of the basic information associated with the job, such as submitter, project, assay, last update, and so on. Details on the software tools used in the analysis of the data (including version information) are also provided – this information is beneficial in ensuring repeatability of analysis and in enabling benchmarking of different software tools. For more advanced users, who may wish to repeat the entire analysis themselves or portions thereof, a log of all processing steps applied to data is also available for download. This log includes parameters used for each of the pipeline applications as well as any output generated, and can be used for troubleshooting in the event of any errors.

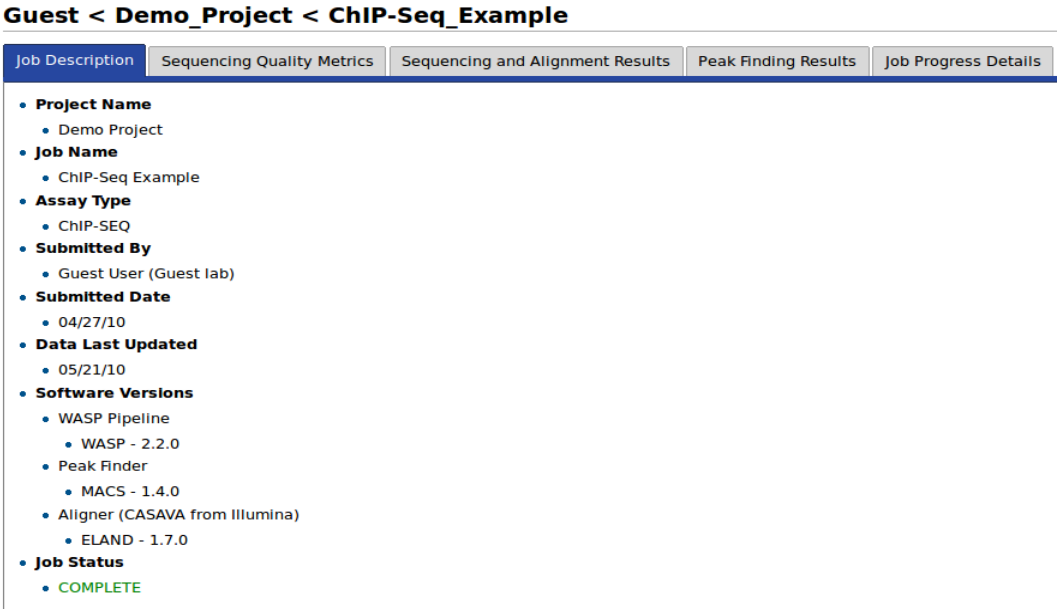

Figure 2.9: Job Description Tab This tab displays the basic information regarding the submission as well as providing details on the software used to process the data and the current job status.

In any biological assay, quality control (QC) is of utmost importance if one is to have any degree of confidence in the results. As sequencing-based assays comprise many different steps and algorithms applied to the data can be quite complex, the WASP system was designed to provide a wealth of easy to understand feedback to users on QC relating to each step of the processing pipeline.

The first section of the Sequencing Quality Metrics tab (shown in Figure 2.10) includes basic sequence statistics such as number of clusters passing the Illumina purity filter, number of sequences

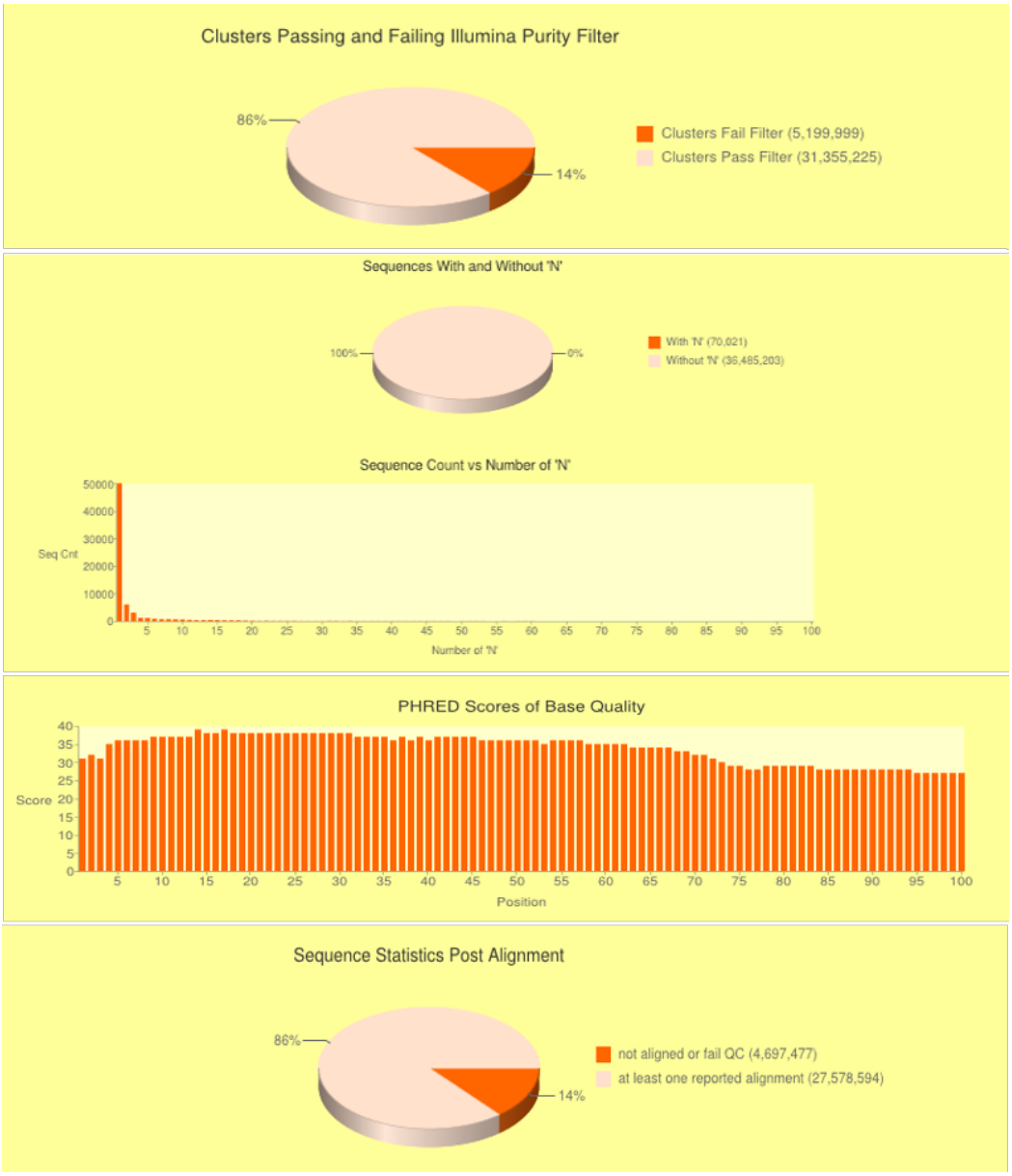

Figure 2.10: Sequencing Quality Metrics The various sections of this tab display basic summary information on the generated reads and can be helpful in troubleshooting. In this example, we see that 86% of the reads pass the Illumina purity filter and that relatively few of those reads contain ambiguous bases. The PHRED quality scores show a slight drop after 75bp but are maintained at a value of above 25 (between 99% and 99.9% accuracy). Almost all of the reads passing filter align to at least one position on the genome.

containing Ns and their distribution, PHRED scores [122] and alignments statistics. This information may be helpful, for example, in determining that a lower than expected percentage of sequences

alinging to the reference genome may be explained by lower quality scores towards the 3' end of the read and that read-trimming may be appropriate.

The sequencing chemistry and processing which happens directly on the instrument are also inherently benchmarkable as Illumina provides guideline ranges for various metrics including: total yield, raw cluster count, clusters passing filter, percentage of phasing and pre-phasing, as well as measures of cycle intensity. While the sequencer generates an XML file containing this data which can be used to generate a table or HTML page, it can be quite difficult to parse in text format given the amount of information. The WASP system instead takes this XML file, extracts the metrics, and then presents them to the user in an easy-to-use 'Googleometer' format (Figure 2.11) generated using the Google Graphs[6] API. This format provides an immediate visual overview of the sequencing run allowing users to quickly identify any areas of concern which may need troubleshooting.

Metric	Read	Result	Uniformity (across tiles)	Notes
Raw Cluster Count	1			Result is 761,567 +/- 101,909 (target is >250,000).
% Clusters Passing Filter (PF)	1			Result is 86.61 +/- 8.39 (target is > 80%).
% Phasing	1			Result is 0.1515. Should be ~0.5% to no more than 1% but in any case, as low as possible.
% Prephasing	1			Result is 0.2618. Should both be ~0.5% to no more than 1% but in any case, as low as possible.
First Cycle Intensity	1			Result is 1,168 +/- 101. Should be >500.
20th Cycle Intensity as % of First	1			Result is 83.15 +/- 1.99. Should be >50%. If too high, suspect relatively low first cycle intensity

Figure 2.11: Run Quality Metrics The WASP system leverages the Google Graphs API to present data on multiple run metrics (as well as expected guideline ranges for these metrics) in an intuitive format.

Finally, we also make use of the third party tools FastQC and FastQ Screen from the Babraham Institute[7] to round out the QC information returned to the user. FastQC includes information on sequence quality, GC content, sequence length distribution, kmer content and any over-represented sequences, while FastQ screen provides the ability to screen a sample of the sequences generated against user-defined libraries. This allows a quick determination if sequences are mapping to the expected organism genome, as well as providing an indication of the presence of any contaminants (Figure 2.12).

The next tab in the results page is the Sequencing and Alignment Results tab (Figure 2.13). This section provides links to download raw FASTQ files as well as SAM and BAM [123] alignment files and BAM index files. These download sections are ordered hierarchically by flow cell ID, sample name, lane, and multiplex index. In cases where PCR duplicates have been removed by the Picard

[6] http://developers.google.com/chart
[7] http://www.bioinformatics.babraham.ac.uk

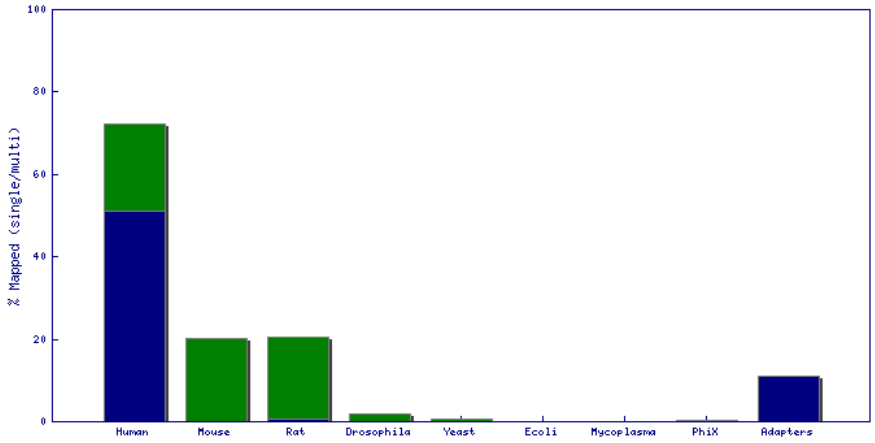

Figure 2.12: FastQ Screen Here, the blue bar represents the percentage of reads mapping only to the current genome while the green bar represents the percentage of reads mapping to both the current genome and also to at least one other genome in the set of libraries. For this analysis 1 million reads are randomly selected from the FASTQ file and are trimmed to 36bp (or 25bp for HELP-tagging or 21bp for miRNA-seq).

Guest < Demo_Project < ChIP-Seq_Example

Job Description | Sequencing Quality Metrics | Sequencing and Alignment Results | Peak Finding Results | Job Progress Details

MD5 checksums for files (often requested when submitting raw data to repositories e.g.GEO or SRA)

Flowcell ID	Lane	Index	Sample Name	Pair	Raw Data Files	Alignment Result
	5	0	input_HD_STR_4w	0	Download Fastq Sequence File Download ELAND-extended Alignment File	***PCR duplicates have been removed by Picard*** Download BAM File Download BAM index File Download Picard Mark Duplicates metrics File Download SAM File Display reads in Genome Browser
61HRLAAXX	6	0	HD_STR_4w	0	Download Fastq Sequence File Download ELAND-extended Alignment File	***PCR duplicates have been removed by Picard*** Download BAM File Download BAM index File Download Picard Mark Duplicates metrics File Download SAM File Display reads in Genome Browser

Job Description | Sequencing Quality Metrics | Sequencing and Alignment Results | Peak Finding Results | Job Progress Details

Sample Name	Sample Type	Flowcell ID	Lane	Result (IP only)	Result (IP vs INPUT)
				Click to download	Click to download
HD_CTX_12w	IP	61HRLAAXX	lane_7_P0_I0	Show peaks in Genome Browser Show wiggle track for IP in Genome Browser Click to download .xls information on called peaks Click to download PDF image of peak model	Show peaks in Genome Browser Show wiggle track for IP in Genome Browser Show wiggle track for INPUT in Genome Browser Click to download .xls information on called peaks Click to download .xls information on negative peaks Click to download PDF image of peak model
Input_HD_STR_4w	Input	61HRLAAXX	lane_5_P0_I0		

Figure 2.13: Sequencing and Alignment and Peak Results Tabs The main sections of the wikipage for return of ChIP-seq results includes the Sequencing and Alignment and Peak Results tabs. These tabs provide links to download or display raw and processed reads as well as the results of any assay-specific downstream analysis provided by the WASP system.

Tools[8] MarkDuplicates function, a link to this duplicates file is also provided. MD5 checksums are made available for raw data files for use in submissions to public repositories such as GEO [124] or the SRA [125]. For ChIP-seq jobs, the final tab contains the results of peak-calling by the MACS [116] software which is automatically invoked as part of the WASP ChIP-seq pipeline. Raw IP and input data as well as called peaks can be displayed as wiggle tracks in WASP's local mirrors of the both the UCSC genome browser [126] and the Broad's Integrative Genomics Viewer (IGV) [127]. Peak information and peak model files can be downloaded as Excel and PDF files respectively.

This section has provided an overview of how an end-user interacts with the WASP system, from registration, through sample submission and return of experimental results. In the next section we examine the system from the point-of-view of core facility personnel.

2.3.3 Core LIMS

Once a submission has been approved and the samples have been physically deposited to the Epigenomics Shared Facility (ESF), core personnel make use of a custom-designed PHP-based LIMS to access the WASP primary datastore and record metadata linking sample information to a sequencing library. The LIMS provides options to create, view, and manage users, labs, samples, jobs, libraries, flow cells, and sequencing runs, as well as perform various billing and administrative functions. Information on core protocols, multiplexing indexes and adaptor sequences can also be accessed. Users are notified on login of any new jobs which have been submitted and not yet processed, and can opt to filter jobs displayed based on new, active, complete, or withdrawn status.

ESF Job List

New Jobs

New Jobs | Active Jobs | Completed Jobs | All Jobs
ESF-Terminated Job | Rejected Jobs | Withdrawn Jobs

JOB ID	JOB NAME ASSAY PLATFORM	SUBMITTED BY EMAIL ADDRESS LAB PI	GRANT FUNDING	PI APPROVAL	DATES & STATUS	ACTIONS
10925 View Wiki Page Job As Textfile Job As Webpage Quote	chipseq2_2 ChIP SEQ ILLUMINA **(HISEQ2500)**	Yuhong Fan yuhong.fan@biology.gatech.edu Bouhassira PI: Eric Bouhassira	Internal User Grant #: 331-005 Awaiting Funding Confirmation	APPROVED 10/06/2013	SUBMITTED: 10/04/2013 STATUS: AWAITING ACCEPTANCE Accept Reject Withdraw Note: Comments will be displayed to user on WASP Comment: Submit	Job Details Sample Detai Library Detail
10924 View Wiki Page Job As Textfile Job As Webpage Quote	chipseq2_1 ChIP SEQ ILLUMINA **(HISEQ2500)**	Yuhong Fan yuhong.fan@biology.gatech.edu Bouhassira PI: Eric Bouhassira	Internal User Grant #: 331-005 Awaiting Funding Confirmation	Awaiting PI Response	SUBMITTED: 10/04/2013 STATUS: AWAITING ACCEPTANCE Accept Reject Withdraw Note: Comments will be displayed to user on WASP Comment: Submit	Job Details Sample Detai Library Detail

Figure 2.14: ESF Jobs Page In this example we see two ChIP-seq jobs submitted on the same day by a single user, one of which has already been approved by the lab PI. Both jobs are currently awaiting funding confirmation based on the grant number provided. The option to provide additional feedback to the user via the WASP system on decisions to accept, reject, or withdraw the job is shown, as are the buttons to view the full job and sample details.

[8] http://picard.sourceforge.net

The standard job view page (Figure 2.14) provides information on job name, assay type, platform, job submitter, PI, and grant funding status and contains links to view the job as a text file or a wikipage. An option is also available to display the automatically generated quote for the job. To the right of each job in the list is a button which moves the user to a more detailed view of the job information or a display of all of the sample metadata associated with this job.

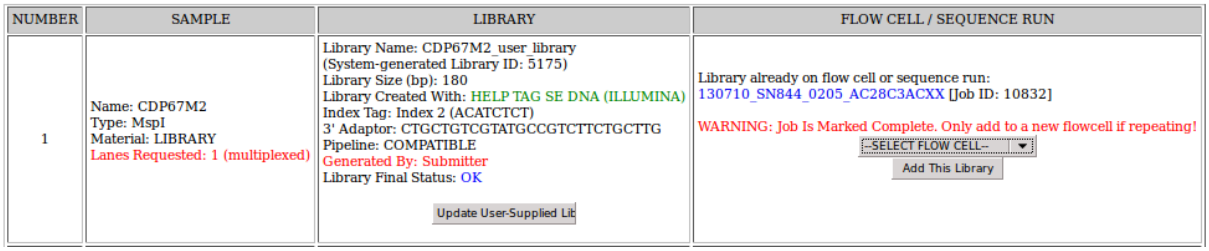

Figure 2.15: ESF Library View Note that, in this example, the sequencing run for the flow cell with which this library is associated is already complete and so a warning is shown beside the option to add this library to a flow cell.

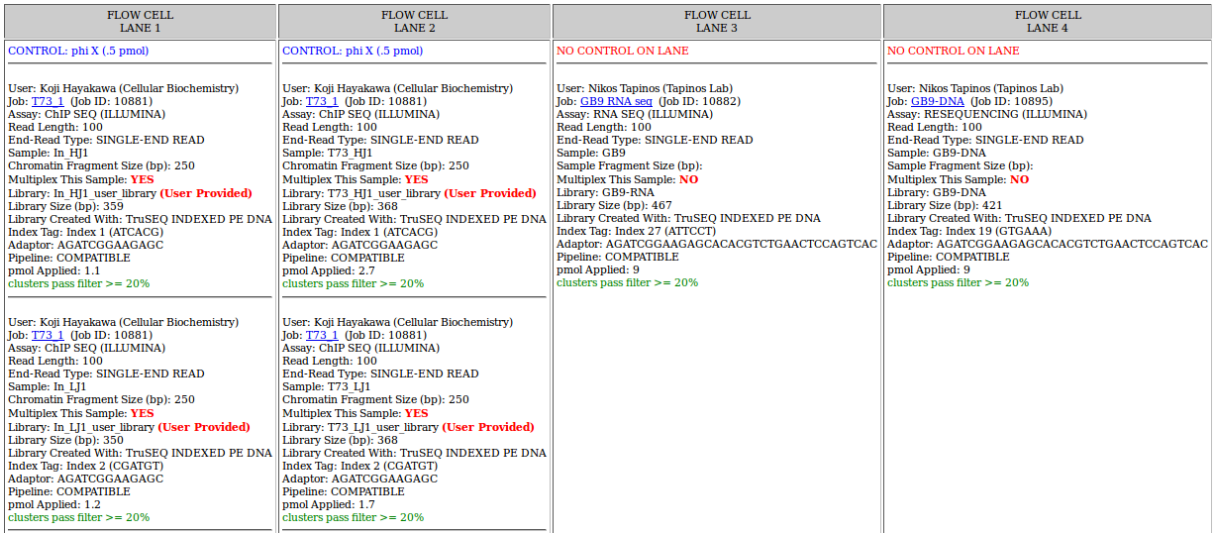

Figure 2.16: ESF Flow Cell View This view shows a physical layout of the flow cell with libraries assigned to specific lanes according to sample multiplexing design. Links are available to job information for each library. In this example, the flow cell has already been run and information regarding the number of clusters passing the purity filter is shown in green. This is another QC step to allow core staff to quickly identify issues with any individual library.

When users submit samples to the facility, they have the option to either allow core staff to generate the necessary sequencing libraries for them for an additional cost, or alternatively to generate their own. Figure 2.15 shows details for a user-supplied library generated using the HELP-tagging single-end protocol, and includes information on library size, 3' adaptor sequence, and multiplex-

ing index used. The functions included in this LIMS view allow core personnel to either modify a user-supplied library, or add this library to an existing flow cell.

The flow cell overview page (not shown) allows users to manage existing flow cells or create new flow cells for any of the supported sequencing platforms, defining parameters such as read length and end-type. An option is also provided to move to a detailed view of the flow cell which displays a lane-by-lane layout and provides information on any libraries assigned to each of the lanes. Figure 2.16 shows an example of this (for readability, only the first four lanes are displayed). We see that this flow cell is designed for a 100bp single-end run and that the first two lanes show multiplexed ChIP-seq samples as well as PhiX spike-in controls.

At this stage in the processing, user, project, job, sample, library, and flow cell information have been linked in the system and all that remains is to link the flow cell ID to a sequencing run. A unique identifier is automatically created for each run in the WASP system by combining run date, machine ID, and flow cell ID. This chain of links from project through to run is shown in Figure 2.17 which provides a partial view of some of the over 40 tables in WASP's main MySQL database used in capturing all of the data and metadata associated with a sequencing job.

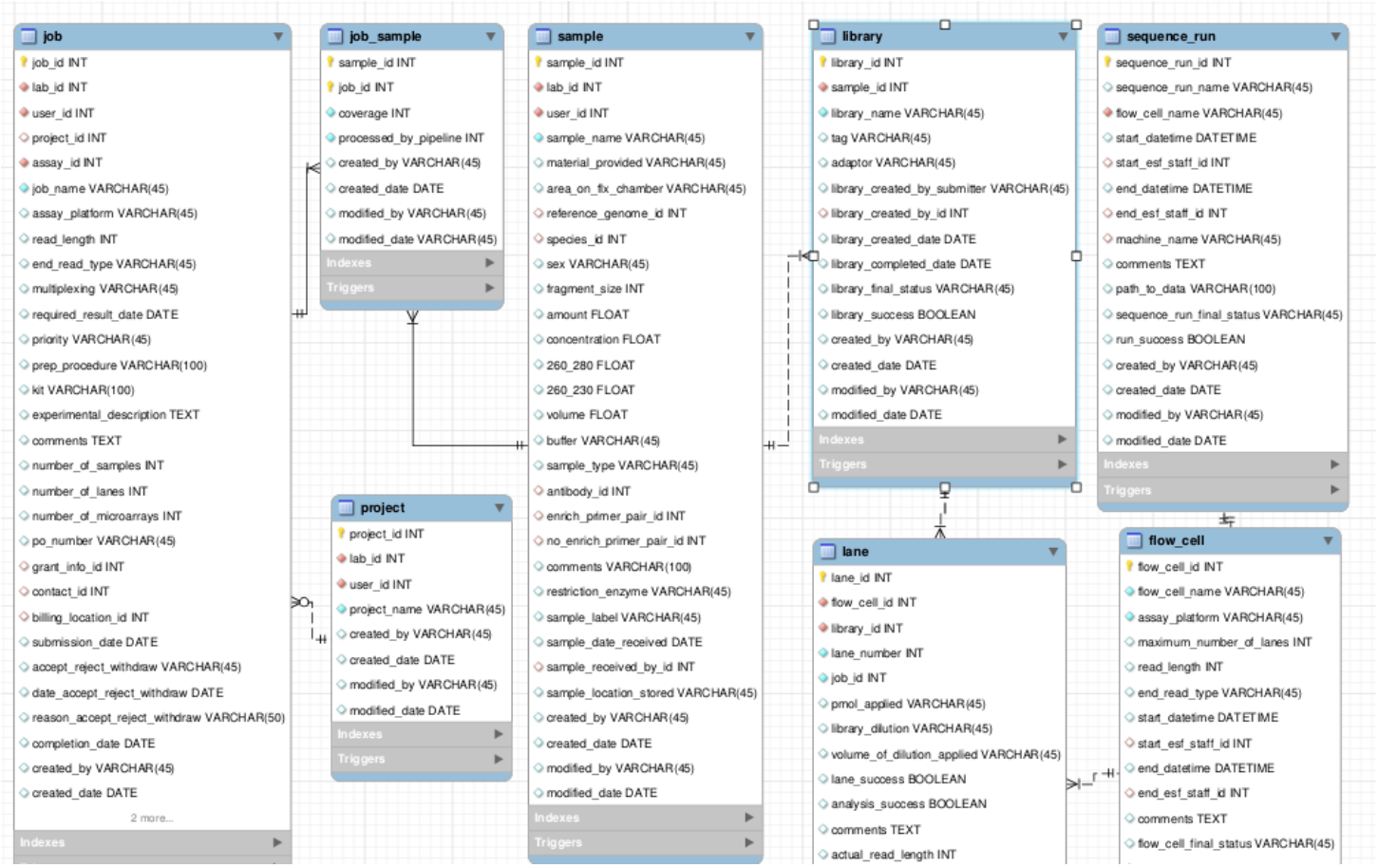

Figure 2.17: WASP Database Shown here are some of the main tables used to capture information about a sequencing run. Each project can consist of multiple jobs, allowing users to logically group related assays. The 'job-sample' table is a join table, used to break the many-to-many relationship between the 'job' and 'sample' tables which arises from the fact that, while a job may have many samples, the same sample can also appear in many jobs. Each sample is related to (potentially) multiple libraries which can be placed on multiple lanes, and a flow cell consists of multiple lanes and is used in a single sequencing run (flow cell stripping and re-use is technically possible but is not practised as part of core standard operating procedures).

From here, the flowcell can be placed in an instrument and sequencing can commence. As soon as the run is complete, the flow of data processing shifts to the backend modules of the WASP system which are described in the next section.

2.3.4 Backend Processing

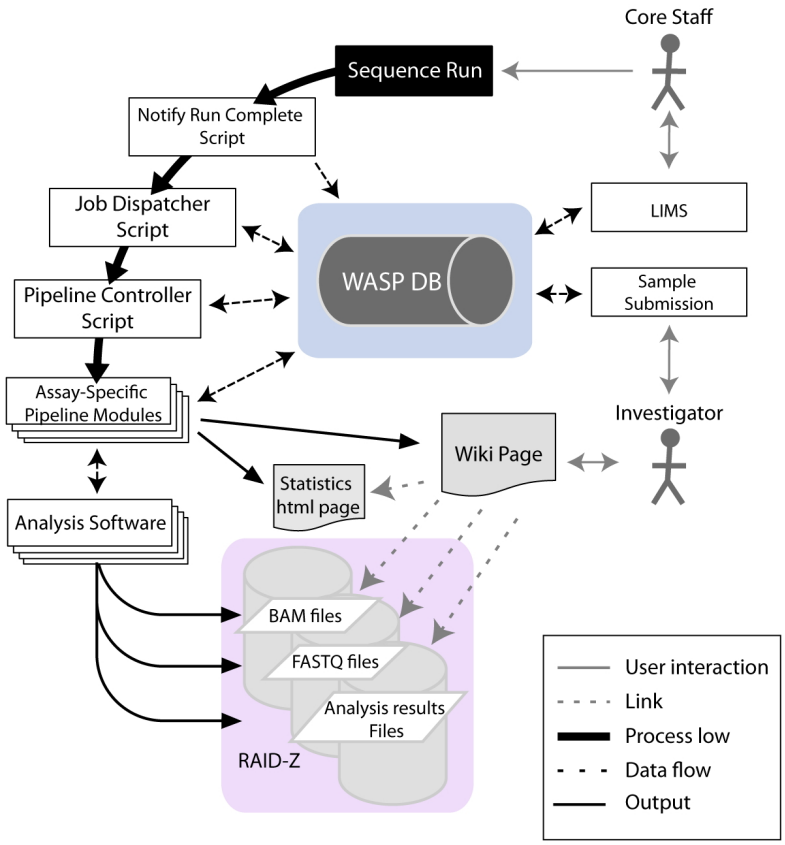

Figure 2.18: WASP Processing Overview Core staff enter information about each sequencing run using a custom built facility LIMS. Once a run is completed, the WASP pipeline controller is automatically invoked and the data is passed to the appropriate assay-specific pipeline for processing. These pipelines, which use community accepted third party tools, also generate appropriate run statistics and update the user wiki page associated with the job with analysis results as they become available. Once the analysis has been completed, the investigator is sent an email with a link to their Wiki page in order to view the sequencing results. (Source: [120])

Once sequencing is complete, raw data files are automatically copied from the instrument to a HPC storage location where the WASP processing scripts can access them (Figure 2.18). These data folders are constantly monitored by a daemonised 'watcher' script, whose job it is to invoke

an instance of the WASP pipeline on any newly-generated data. The watcher script first parses the run name to extract the flow cell ID and then adds information to the sequence run table in the database indicating the path to the data. A list of job IDs associated with this flow cell is then pulled from the database and each job is passed to an instance of the pipeline controller script. The main WASP backend processing is handled by a set of assay-specific Perl modules, although several additional programming and scripting languages (including Python, AWK, R, and C++) are used for certain tasks as necessary. With common functionality being derived from inherited base classes, these Perl modules are largely self-contained making them easily modifiable. The pipeline controller script uses the job information provided to ascertain the type of assay for a particular job and then calls the appropriate assay module to further process the data. Currently, ChIP-seq, RNA-seq, microRNA-seq, and HELP-tagging are supported, and an overview of the processing performed by each of their associated modules is given in Figure 2.19.

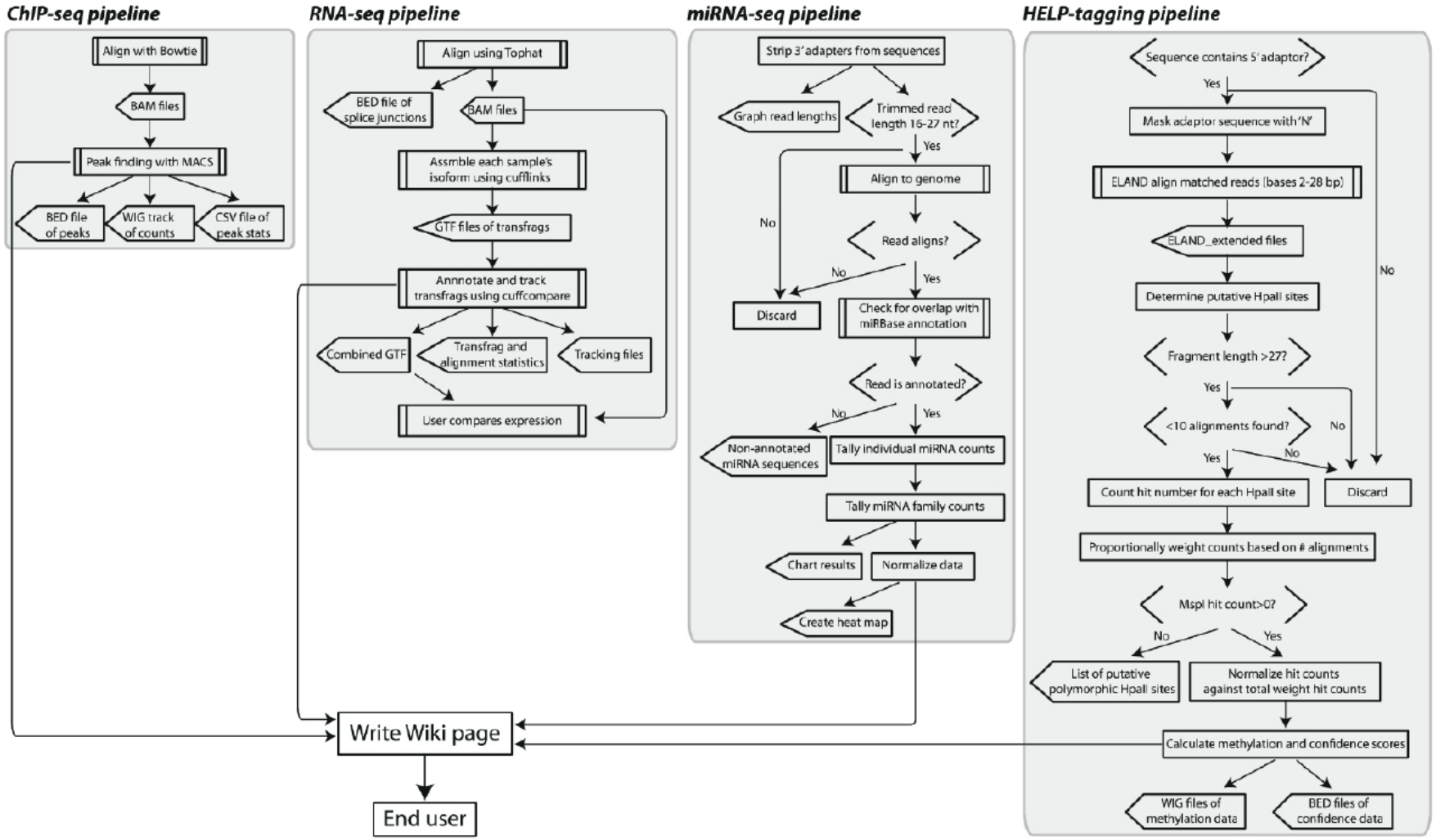

Figure 2.19: WASP Pipelines Overview Processing steps are shown for the four main assays supported by the WASP system. The end result of all pipelines (including any non-pipeline compatible jobs) is the writing of the results wikipage and automated email to the end user informing them that their analysis is complete. (Source: [120])

As previously stated, WASP was initially designed to work with the Illumina platform requiring that the first step in post-processing be to convert the proprietary QSEQ format files to Sanger FASTQ format, demultiplexing as necessary. In the case of ChIP-seq jobs, FASTQ files are then aligned to the reference genome specified during sample submission to produce SAM and BAM [123] format files, after which peaks are called. The WASP system makes use of standard community-accepted third party tools for sample processing, examples of which include the previously mentioned BWA [103] or Bowtie [106] for alignment, MACS [116] for ChIP-seq peak-calling, and Tophat [128] and Cufflinks [129] for processing of RNA-seq data[9]. The modular design of the system facilitates the relative ease of swapping these tools as desired.

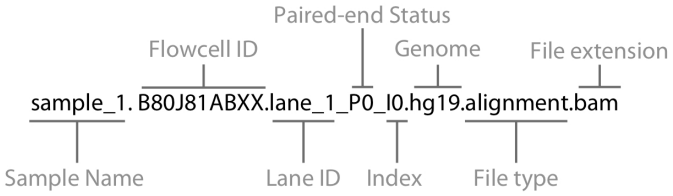

Figure 2.20: File Naming For paired-end status: 'P0' indicates unpaired, 'P1' indicates first mate, and 'P2' indicates second mate. Multiplexed index of '0' indicates sample is not being multiplexed. (Source: [120])

We have previously mentioned the naming convention of 'date-machine-flowcell' for individual sequencing runs. WASP also uses a naming convention for results files generated by the various processing pipelines. These human readable names are designed to be both informative and easily parsed by pipeline scripts, and consist of: sample name, flowcell ID, lane ID, paired-end status, index used for multiplexing, reference genome, filetype, and file extension. An example of this is shown in Figure 2.20. The WASP system also maintains a strict hierarchical directory structure for these results files, with each job belonging to a particular project, each project belonging to an individual WASP user, and each user belonging to a labspace defined by the WASP ID of the lab's PI (as indicated during registration). Within a job folder, files are further divided into 'raw', 'processed', and 'analysed' results folders with data organised accordingly.

2.4 Discussion

Related Tools

At the time the WASP system was being developed (2009-2010), several other solutions to the informatics challenges of NGS were also in the early stages of development. The LIMS component was

[9]The HELP-tagging assay was designed in-house necessitating a custom-built pipeline.

being addressed by both open source efforts such as The Genome Analysis Centre's MISO (Managing Information for Sequencing Operations)[10], as well as commercial ventures such as Bioteam's WikiLIMS[11], a MediaWiki instance with semantic extensions. Workflow engines such as Taverna [130] and Conveyor [131] were providing flexible ways to define, execute, and share analysis pipelines, making use of third party tools available through REST- and SOAP-based web services, while other platforms such as geWorkbench [132], KNIME [133], and Galaxy [134] focussed on providing a suite of data processing, analysis, and integration tools. Notably however, none of these efforts at that time provided the integrated sample submission, LIMS functionality, automated analysis pipelines, and visualisation capabilities which were included as part of WASP's end-to-end design.

The WASP System

Once WASP was in full production mode, the system was presented at various national and international meetings, garnering widespread interest from other institutions where sequencing facilities were being established. These new centers were struggling with some of same fundamental issues which had initially led to the development of the WASP system, and were eager for a turn-key solution to their data-capture, -management, and -analysis problems. The issue however, was that the WASP system had been developed specifically to meet the needs of the Einstein Center for Epigenomics and was therefore not readily customisable to the requirements or infrastructure of other institutions, or indeed to other sequencing platforms, having been primarily developed with Illumina technologies in mind. The system was also only modularised to a certain extent, having grown somewhat organically during early development in response to both frequent changes in Illumina's algorithms, and file formats, as well as to demand for quickly supporting an increasing range of sequencing assays. It was therefore decided to re-implement WASP in a highly modular and extensible way, leveraging mature, enterprise-level technologies to produce a system that could be easily configured by any interested institution or facility to their own specific requirements. This new flexible design also aimed to provide individual researchers with greater control over the analytical pipelines, allowing them to choose tools and parameters to suit their own needs.

The programming paradigm for this new system, termed 'The WASP System' (as opposed to the original 'WASP'), is based around the Java/J2EE Spring framework. This application framework, developed by SpringSource[12] contains multiple Spring Projects, each of which provides support for different aspects of an enterprise-level infrastructure. The core Spring framework supports (amongst others): an inversion of control (IoC) programming model (via dependency injection) which allows abstraction of application logic into decoupled, re-usable modules, transaction management and aspect-oriented programming (AOP), web applications using the model-view-controller (MVC) approach, data access and object-relational mapping (ORM) using Java Database Connectivity (JDBC) and integration with ORM libraries such as Hibernate, message passing via the Java Messaging Service (JMS), and unit and integration testing using, for example, the popular JUnit

[10] http://www.tgac.ac.uk/miso
[11] http://bioteam.net
[12] http://spring.io

module. Examples of WASP re-development using Spring-based technologies includes: 1) the AJAX submission forms being replaced by a combination of Java Server Pages (JSP) and Spring Web Flow (SWF), which dynamically build sections of a web page according to XML-based flow definition files, 2) MediaWiki access extensions and custom-written Perl scripts being replaced by the much more powerful Spring Security, which allows highly-customisable authentication and role-based access definitions, protecting against vulnerabilities such as session hijacking and clickjacking, and 3) Spring Batch replacing the complex scripted controllers for the assay-specific pipelines. Spring Batch was designed to simplify the management and life-cycle of high-volume batch jobs or workflows and provides functionality and application programming interfaces (APIs) to start, stop, and gracefully restart jobs on interruption. It also includes job tracking and statistics as well as web-based job management capabilities.

To further increase the modularity in the system, each of the components (database, web front-end, security, pipelines, and so on) are deployed as a separate Spring application bundle on an Eclipse Virgo OSGi[13] server. The OSGi model is designed for module-based Java applications and provides a more robust production environment in which each of the components can be developed, updated, and, if necessary, shutdown independently, while allowing the other component to continue functioning.

In addition to streamlining and modularising the existing functionality of WASP, the new WASP System has two further key design goals (as shown in Figure 2.21). The first of these stems not only from the desire to create an extensible system, capable of integrating diverse datasets and facilitating a systems approach, but also from the recognition that no one institution will likely be able to develop a solution to all of the sequencing and analytical challenges faced by the genomics and epigenomics communities. To that end, a 'nurtured open-source distributed development environment' was envisaged, with the aim to both encourage adoption of the new system, and, more importantly, to dramatically increase the number of programmers actively participating in the development process. The WASP System therefore comprises two components – the first is the core system, which, for stability purposes, has a restricted codebase which will continue to be developed by a small group of key programmers. This core however includes a plugin-handler which hosts the second system component – a collection of third-party plugins developed by programmers from geographically diverse institutions. The idea behind this design is that, while development of some plugins may serve specific individual institutional needs and be integrated only into local installations of the WASP System, other plugins will likely address challenges shared by many of the participating institutions and will be made available to the entire WASP System community as part of the shared plugin collection resource. In order to ensure that any plugins developed are compatible with the core system (including any current pre-release versions), plugin developers will be provided with template plugins and documentation, and will have access to the continuous integration and testing servers currently hosted at Einstein. It is important to note that these plugins are not solely limited to genomic or epigenomic data – proteomic, metabolomic, and any other form of data can be handled

[13] http://www.eclipse.org/virgo/

provided that the appropriate Spring Web Flows for dynamic metadata capture, JSP pages for data display, and Spring Batch XML definition files detailing appropriate processing steps are defined.

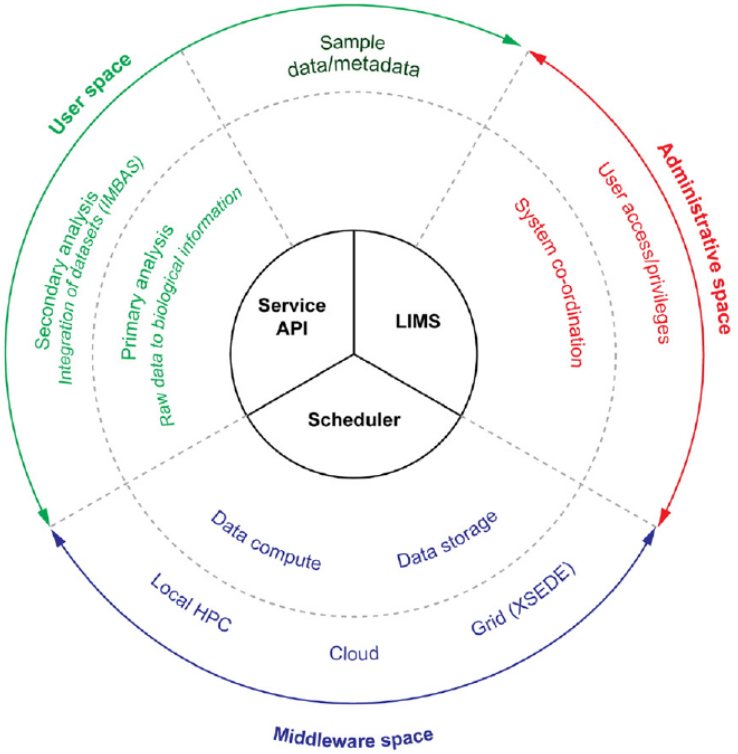

Figure 2.21: The WASP System This figure shows the main components of the new design, incorporating all of the functionalitiy of the original WASP, as well as new functionality relating to data integration and the system's role as middleware for grid- and cloud-based HPC resources. (Source: [120])

The second new design goal addresses the fact that, as well as having different software ecosystems, different institutions will likely have access to vastly different hardware and HPC resources. Many facilities may not have bespoke local compute clusters and may instead rely on grid or cloud-based resources to provide scalable, on-demand computing for handling their sequencing data. With compute and storage costs in the cloud decreasing, this option may be particularly attractive to smaller institutions or core facilities who may prefer not to have the overhead associated with sourcing, maintaining, upgrading, and replacing expensive HPC systems and hiring personnel specifically for their administration. The desire to increase the system's flexibility in this regard also coincided with Einstein's increased interaction with the NSF-funded XSEDE[14] (Extreme Science and Engineering Discovery Environment) national grid resource in the form of the Einstein Genome Gateway – an XSEDE-supported portal for providing shared access to grid-enabled genomics and epigenomics tools. The WASP System therefore leverage's existing software such as Cloud tools[15] and the Crux

[14] http://www.xsede.org
[15] http://code.google.com/p/cloudtools/

Toolkit[16] to enable it to act as middleware for cloud and grid resources [135, 136].

The new WASP System is, at the time of writing, currently being tested and customised at three partner institutions – Memorial Sloan-Kettering Cancer Center, New York University, and the University of California San Diego. Once this first phase of testing has been completed, it will be rolled out to up to twenty further national and international academic and commercial partners, including the newly-established New York Genome Center[17] in a project dubbed the 'WASP Swarm'.

[16] http://confluence.globus.org
[17] http://www.nygenome.org

CHAPTER 3

Secondary Analysis of ChIP-Seq Data

The ChIP-chip datasets used in this chapter to test the ChIPSOM algorithm were generated in the labs of Prof. Jonathan Licht, Northwestern University (WT1 dataset) and Dr. Anton Krumm, University of Washington (CTCF dataset). All algorithm development and testing was carried out by the candidate.

3.1 The Motif Finding Problem

A common task in the analysis of genomic sequence is the identification of short, recurring patterns, known as motifs, which, due to their over-representation, are assumed to have some biological significance [137]. While sequence motifs can be used to model, for example, restriction enzyme recognition sites, ribosome binding sites, splice sites, and miRNA binding sites, in this section we focus on the use of motifs for modelling transcription factor binding sites (TFBSs). Motif discovery is an important step in the secondary analysis of ChIP-seq data for several reasons. Firstly, it confirms the presence of a *bone fide* binding site within a ChIP-seq peak adding confidence to the called peak. Secondly, despite the fact that binding sites are quite small, ChIP-seq peaks can range in the order of hundreds of base pairs; motif identification can therefore help to increase the resolution of the binding site annotation. Thirdly, not all binding sites are created equal – variations in binding site sequence can result in a spectrum of binding affinities, potentially modulating the regulatory effect of the associated transcription factor. Using *de novo* motif discovery can help to uncover the different modalities of a binding site and show any divergence from known canonical motifs. In this section we discuss the different ways in which TFBS motifs can be represented and describe some common approaches to their *de novo* identification.

3.1.1 Motif Representation

The response elements (REs) to which transcription factors bind are short, degenerate sequences, usually in the order of 6–32bp in length [138]. The sequence degeneracy reflects the fact that the interaction between protein and DNA is highly dependent on structure, and nucleotides not involved in binding are generally not subject to the same selective pressures as those that are [139]. Transcription factors may therefore recognise and bind to non-functional elements located throughout the genome (spurious binding), creating a need for increased protein production and placing extra demands for energy expenditure on a cell. While one may think that such spurious sites should therefore be under purifying selection, recent research suggests that conserved chromatin context can play a role in relaxing selection against these sites [140]. It has also been suggested that the clustering of functional binding sites into regulatory modules helps to target binding to functional elements thereby reducing the effects of spurious binding [141].

Symbol	Description	Bases
A	**A**denine	A
C	**C**ytosine	C
G	**G**uanine	G
T	**T**hymine	T
M	a**M**ino	A and C
R	pu**R**ine	A and G
W	**W**eak interactions	A and T
S	**S**trong interaction	C and G
Y	p**Y**rimidine	C and T
K	**K**eto	G and T
V	Not-T (**V** follows T)	A and C and G
H	Not-G (**H** follows G)	A and C and T
D	Not-C (**D** follows C)	A and G and T
B	Not-A (**B** follows A)	C and G and T
N	a**N**y	A and C and G and T

Figure 3.1: IUPAC Codes IUPAC codes for representing consensus sequences are shown. Each symbol represents the bases which can occur in any given position within a consensus sequence.

By aligning multiple experimentally-derived binding sites into a motif, we can capture the general binding preference of a particular transcription factor. One simple way to do this is through the use of a consensus sequence. Consensus representation usually takes one of two formats. The first format involves the use of regular expressions using the standard DNA alphabet ($b \in$ {A,C,G,T}); an example of this is a sequence such as A{G}CC[CG]T, where a single letter represents a fully conserved base, {G} indicates any base but G can occupy this position, and [CG] indicates that

either a C or a G occur in the sequence. The second method extends this alphabet with IUPAC[1] codes, adding letters which represent the possibility of two, three, or any of the four nucleotides occurring in each position in the motif (Figure 3.1). While consensus sequence representations are relatively straightforward, they do however result in loss of information – there is no way to tell, for example, if one of the nucleotides in a two-base consensus occurs more frequently than the other.

A more comprehensive approach which does not suffer from this problem is the use of a position specific scoring matrix (PSSM). A PSSM is a raw count of the number of times each nucleotide occurs in each position in the motif, this results in a 4 x ℓ matrix, where ℓ is the motif length. Normalizing for the number of sites used to create the matrix produces a position specific frequency matrix (PSFM), in which f_{bi} is the probability of observing nucleotide b in position i and $\sum_b f_{bi} = 1$ (i ϵ {1,2,...,ℓ}). Extending this representation to incorporate background frequencies (thus allowing for consideration of different nucleotide frequencies in different organisms being studied [137]) produces a position weight matrix or PWM. Here, W_{bi} is the log-odds for base b in position i of the motif, and p_b represents the probability of base b occurring in a background model:

$$W_{bi} = log_2(\frac{f_{bi}}{p_b}). \tag{3.1}$$

While providing a more powerful way to capture motif information, matrix-based approaches do have some associated problems. The first of these is that, when constructing matrices from a small number of samples, there is a possibility that a particular nucleotide may not occur in the input sequences. This results in an undefined log-odds value ($log_2(f_{bi}/p_b) = log_2(0)$). In order to correct for this occurrence, small pseudocounts are usually added to the frequency count at each position in the motif [142]. A further limitation of this type of representation is the assumption that all positions in the motif are independent, which is not the case. This has been demonstrated by, for example, Bulyk et al. [143], who use a microarray-based approach to show that the binding affinities for mutant and wild-type early growth response 1 (EGR-1) zinc finger transcription factors are highly dependent on inter-nucleotide effects. PSSMs and their derivatives also do not accommodate gaps, which are found in some motifs which consist of half-sites separated by variable length spacer regions [144]. There are several approaches which aim to address these issues either using simple pairwise nucleotide dependencies [145], or by employing more advanced models. The authors in [146], for example, use permuted variable length Markov models (PVLMMs) to capture dependencies among nucleotide positions and demonstrate the application of these models to the detection of both splice sites and transcription factor binding sites. More complex models such as these, however, usually require more biological knowledge for their construction and offer only marginal improvement in specificity; for this reason, matrix-based approaches are still the currently preferred method for modelling binding site motifs.

A graphical motif representation was developed in [147]. These 'sequence logos' show each column

[1]IUPAC – http://www.iupac.org/

of the PSSM as a stack of letters, and characterises the conservation of the individual nucleotides at each position in a motif in terms of information content; each letter's height in the logo is proportional to its observed frequency (see Figure 3.2). They can be constructed using the following equation:

$$IC_i = 2 + \sum_b f_{bi}\, log_2(f_{bi}), \tag{3.2}$$

where, as before, f_{bi} is the probability of observing nucleotide b in position i, $b \epsilon \{\text{A,C,G,T}\}$. Perfectly conserved positions in this model will contain 2 bits of information, two-base degeneracies will convey 1 bit of information, and positions where all four nucleotides can occur are non-informative. It has been shown that a motif's information content is directly related to both its expected frequency in a random DNA sequence [137] and the average binding energy of all sites used in its construction [148]. Note that equation (3.2) assumes equal nucleotide probabilities, this model can be generalised in the same way as a PWM by using the relative entropy of the observed nucleotide frequencies with respect to background composition of different organisms:

$$IC_i = -\sum_b f_{bi}\, log_2(\frac{f_{bi}}{p_b}). \tag{3.3}$$

An example showing the different motif representations discussed above for a set of four short sequences is shown in Figure 3.2.

Once we have a representation of a motif, we can use it to scan the promoter region of our gene(s) of interest to determine whether or not a particular transcription factor is likely to bind there. If we have not identified our own motif *ab initio* (a topic which we will address in the next section), there exist many databases of previously determined binding motifs which can be used for this task. TRANSFAC [149] and JASPAR [150], for example, contain PSSMs for many eukaryotic organisms, while PRODORIC [151] holds binding data for prokaryotes. Species-specific motif databases also exist, examples being SCPD [152] and RegulonDB [153], which contain information on *S. cerevisiae* and *E. coli* TFs respectively. Programs such as MatInspector [154] use PWMs from these databases to search for binding sites in promoters by sliding a weight matrix along an input sequence and scoring the similarity between the motif and sequence at each starting position using an equation such as:

$$S_x = \sum_b \sum_i x_{bi}\, W_{bi}, \tag{3.4}$$

where x_{bi} is equal to 1 if base b occurs at position i in the motif and is equal to 0 if it does not.

SEQ 1	A	C	C	T	G
SEQ 2	A	T	G	T	A
SEQ 3	A	G	G	T	G
SEQ 4	A	A	C	T	G

PSSM					
A	4	1	0	0	1
C	0	1	2	0	0
G	0	1	2	0	3
T	0	1	0	4	0

CONSENSUS	A	N	S	T	R

PSFM					
A	1	0.25	0	0	0.25
C	0	0.25	0.5	0	0
G	0	0.25	0.5	0	0.75
T	0	0.25	0	1	0

PWM					
A	2	-2	0	0	-2
C	0	-2	-1	0	0
G	0	-2	-1	0	-0.42
T	0	-2	0	2	0

LOGO

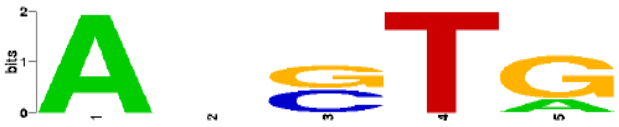

Figure 3.2: Motif Representation The topmost panel shows the multiple alignment for a set of four binding sites. The PSSM represents the raw counts of the individual bases within this alignment while the PSFM is simply the normalised count matrix. The consensus sequence uses IUPAC codes from Figure 3.1 to represent the alignment but results in loss of informaition – notice that the last base in the consensus, R, gives no indication as to the relative frequency of G to A in this position, despite the fact that the occurence of a G is much more likely. The PWM is calculated as the log_2 ratio of the observed to background frequencies, and in this case assumes equal background nucleotide frequencies (0.25). Finally, the sequence logo for the alignment is a graphical representation of the conservation at each position with each letter scaled based on its frequency.

Positions which score above a certain threshold are tagged as the start of likely binding sites. This pattern matching approach to identifying binding sites can, however, suffer from large numbers of false positives, and considerable effort must be dedicated to optimizing thresholds in order to minimise these [155]. One approach used in [156] is to simultaneously optimise both the weight matrix and threshold so that all known binding sites are identified with as few other sites included as possible.

Before we can use motifs in this type of pattern matching, we first need a way to derive the binding model from appropriate input sequences. The original motivation for *de novo* motif-finding problem was expressed in terms of DNA microarrays used for measuring gene expression levels. The rationale is that after clustering the expression profiles, genes which are potentially co-regulated (and therefore likely bound by the same transcription factor) are grouped together. By searching for statistically overrepresented sequences in the promoter regions of these gene clusters, it is possible to identify novel motifs. Because of the nature of TF binding sites, the motif discovery process must attempt to address several difficult issues. Firstly, there is no way to know *a priori* what length the motif is. Similarly, there is no way to know if a given transcription factor binds only once in each promoter (or not at all), or if some of the promoters contain multiple instances of the same site. It is also likely, as previously mentioned, that a promoter region will contain binding sites for many different transcription factors which act in a coordinated manner to regulate expression. The next section looks at some of the different ways in which this problem can be tackled.

3.1.2 Traditional Approaches to *de novo* Motif Finding

Perhaps the simplest approach to determining over-representation is the use of enumerative or word-counting methods like that of YMF [157]. With this approach all ℓ-*mers*, or subsequences of a particular length, ℓ, are enumerated, and the frequency of their occurrences in the input sequences are calculated. Over-representation is then measured by comparing observed to expected frequency, with the expected frequency of a sequence $Seq = b_1,\ b_2\ ,..., b_\ell$ calculated as either:

$$Pr(Seq) = \prod_{i=1}^{\ell} p_{b_i}, \tag{3.5}$$

where p_b is the probability of finding base b in the input sequences, or by using a Markov chain of order m, where a base's probability is calculated based on the previous m bases

$$\begin{aligned} &Pr(B_n = b_n \mid B_{n-1} = b_{n-1},\ B_{n-2} = b_{n-2}\ ,...,\ B_1 = b_1) = \\ &\quad Pr(B_n = b_n \mid B_{n-1} = b_{n-1},\ B_{n-2} = b_{n-2}\ ,...,\ B_{n-m} = b_{n-m})\ ,\ (n > m). \end{aligned} \tag{3.6}$$

Markov models are a useful tool for sequence analysis since they can capture more subtle characteristics of nucleotide frequency than independent nucleotide counts [158]. While an exhaustive search is guaranteed to find existing motifs, an obvious disadvantage of this approach is that for each motif length ℓ, there are 4^ℓ candidate strings to evaluate resulting in exponential increases in search time as the motif length increases [159]. The search becomes further complicated depending on the number of mismatches allowed when comparing to the consensus sequence, and whether or not gaps are permitted. Some variants on this approach try to circumvent these limitations and

improve compute-time using either suffix trees to pre-index the input sequences (Weeder [160]), or by using dictionary-based search such as MobyDick [161].

While early motif discovery tools used enumerative methods, most modern techniques are based on probabilistic approaches. One of the most popular of these is the expectation maximisation (EM) [162] approach used by MEME [163]. MEME uses expectation maximisation to fit a two-component mixture model to sequence data, where component one represents the motif and component two is the background. The E-step uses a PWM generated from random initial positions to calculate the expected log-likelihood over the latent variable (whether a sub-sequence has been generated from the motif or background model), while the M-step is used to compute parameters which maximise the expected log-likelihood from the E-step (i.e. finding the locations in each sequence which align maximally to the current motif). These parameters are then used in the next E-step, and the algorithm proceeds in this iterative manner until convergence. By probabilistically masking sequences for previously found motifs, MEME can identify multiple distinct motifs within a single data set. While MEME may represent an improvement over naïve word-based approaches, it does suffer from problems of its own. Like all motif-finders there are many associated parameters (probable motif length, number of expected occurrences per sequence and so on), but more importantly, this approach can suffer from dependence on initial conditions with no guarantee that it will converge to a global rather than local maximum.

Gibbs sampling [164] is another popular probabilistic approach used by tools such as AlignACE [165], info-gibbs [166], and Motif Sampler [167]. It is similar to MEME given that it uses a combination of EM and simulated annealing, but is less likely to converge to a local maximum due to its leave-one-out sampling strategy (although this is still not guaranteed). With this approach a PWM is initially constructed from random starting positions in each of the input sequences except one (S_x). During each iteration, all sub-sequences of S_x are scored using the PWM to determine the most likely binding site; this position is then used in the construction a new motif, while another sequence, $S_{x'}$, is removed and scored. This process is repeated until the positions in each sequence do not change. One clear disadvantage of this approach is that a model of one motif per sequence is assumed.

While some tools, such as MDScan [168], aim to combine the advantages of both word enumeration and PWM-based approaches, seeking a balance between exhaustive search and rigorous statistical modelling, others such as ConSite [169], rVista [170], and TOUCAN [171] use phylogenetic information to bias the search toward evolutionary conserved regions which may be more likely to contain functional binding sites. This may seem reasonable given that several studies have previously shown that the occurrence of mutations in binding sites happens at a rate two to three times lower than that expected for functionally neutral mutations [172, 139], many other studies, however, have shown evidence for widespread turnover of transcription factor binding sites [173]. Some studies, for example, indicate up to 94% conservation in binding sites for individual proteins between primates [174], while others report that 32-40% of human functional sites are not functional in rodents [173], or demonstrate highly divergent binding profiles even across closely related species of yeast [175]. This apparent contradiction may stem from the fact that regulatory sequences can be shuffled and

re-ordered yet still remain functional, making detection of conservation difficult. We must also consider that not all conserved sites may actually be functionally active. The benefits of incorporating phylogenetic information on binding sites are therefore somewhat difficult to assess – while possibly helpful in reducing the 'noise' inherent in motif finding and focusing the search, it may not only disregard functionally important sites, but also overestimate the importance of non-relevant ones as well. A more promising approach may be to make use of the fact that functional binding sites are usually clustered into *cis*-regulatory modules (CRMs).

There are two classes of CRM detection algorithm. The first class starts with individual binding sites and then attempts to build higher order structures and patterns. The second aims for fully *de novo* module prediction based on hierarchical mixture models and Bayesian inference to maximise the joint probability distribution associated with multiple TF binding [176]. Algorithms of both class may seek to exploit prior information on TF binding – the Ahab tool [177], for example, was used with binding motifs known to be involved in *Drosophila* segmentation to predict 146 regulatory modules. Sinha et al. [178] also exploit prior knowledge of *Drosophila* TFs which are known to co-locate, and use a hidden Markov model (HMM) coupled with EM for parameter estimation. With this approach, different states in the Markov model can be used to represent different motif models (or a background model), emitting nucleotides with probabilities corresponding to the observed frequencies in the input sequences. The distances between motifs within a module as well as the expected module length can then be captured by the state transition probabilities. Acquiring prior biological knowledge of positionally correlated motifs is unfortunately however not always straightforward; while some CRMs have strict rules about the order and spacing or binding sites [179], others are less stringent resulting in a combinatorial problem of a much higher order [180]. Several algorithms also exist which combine phylogenetic information with CRM detection; ModuleSearcher [181], for example, biases its search for CRMs towards regions which are conserved between human and mouse; we have, however, already commented on the potential problems with such approaches.

Since different motif finders may sometimes produce different results, perhaps the most promising approach is that of pipelines such as TAMO [182], which executes a combined analysis using the AlignACE, MDScan, MEME, and Weeder programs, performs statistical testing on the output of each, and then clusters the significant motifs. This type of consensus approach, by combining the best hits from several motif finders (and several different approaches to motif finding), is far more likely to succeed in identifying high confidence sites while simultaneously minimizing false positives than any one motif finder alone. An extension of this approach, WebMOTIFS [183], combines TAMO with Bayesian analysis, incorporating prior knowledge about likely motifs.

Finally, it is worth mentioning that one potential limitation of all of these approaches to both individual motif and CRM detection is that, even when incorporating phylogenetic information and accounting for the clustering of TFBSs in regulatory regions, neither protein-protein interactions nor chromatin structure are considered. We have already indicated both a), that many transcription factors bind as heterodimers and that the binding of one TF may serve to recruit or stabilise the binding of another, and b), that local chromatin context (including DNA looping) plays an important role in regulation; it remains to be seen if any algorithms can successfully integrate this kind of prior

knowledge to improve the discovery process. In the next section we discuss an alternative approach to *de novo* motif finding based on the Self-Organizing Map.

3.2 ChIPSOM

The self-organizing map (SOM) [184, 90] is a neural network architecture commonly used in exploratory data analysis to perform unsupervised clustering and visualisation of high-dimensional data. Since its initial description by Kohonen [184], it has been cited thousands of times [185, 186], having been applied to a wide range of problems including datamining [187], image compression [188], and machine vision [189], as well as being used for the identification of patterns in data from fields as diverse as geography [190], finance [191], and biology [192]. In this section we detail the operation of the SOM, outline some related techniques, and describe the SOMBRERO motif finding algorithm previously developed in our lab. We then outline some of the limitations with the original algorithm, provide details on ChIPSOM, a modified version designed to allow processing and visualisation of genome-wide data and then introduce the more fundamental problem of motif redundancy which required the development of a novel post-processing algorithm.

3.2.1 Introduction to SOMs

A SOM consists of a lattice of output nodes onto which the input data is mapped; these nodes are usually arranged in a two-dimensional rectangular or hexagonal shape, although neither the dimensionality nor the shape of the lattice are restricted [193]. Each node i in the lattice contains a parametric model vector (or weight vector) of length n, where n is the number of dimensions in the input space ($m_i = [\mu_1, ..., \mu_n]^T \epsilon \Re^n$). These models can either be randomly initialised, or can be based on for example, the eigenvectors from a principal components analysis (PCA) of the data, with the latter approach resulting in reduced training time. During training (depicted in Figure 3.3), vector quantisation is performed in a two-step process. In the competition phase, an input vector $x = [\xi_1, ..., \xi_n]^T \epsilon \Re^n$ is selected from the training set and the node model it most closely matches is determined. In the learning phase, the matching node previously identified, as well as its neighbours (discussed further in the next section), are updated so that they becomes more similar to the input vector. This process is repeated for each training sample over many iterations eventually producing a similarity graph where the nodes provide a discrete approximation of the distribution of training samples. Note that although we describe the operation of the SOM here in terms of vectors, it is equally possible to use any non-vectorial data (symbol strings for example), provided a suitable metric for comparison (such as edit distance) is used [194].

Determining the best matching unit (BMU) or 'winning' node requires some measure of distance between the input and model vectors, commonly chosen metrics include the Euclidean distance, dot product, Minkowski metric, and Mahalanobis distance [90]. Here we'll assume the Euclidean distance, given by:

$$d(x, m) = || \, x - m \, || = \sqrt{\sum_{i=1}^{n} (\xi_i - \mu_i)^2}. \tag{3.7}$$

The winning node, c, is then selected according to:

$$c = \arg\min_i \{d(x, m_i)\}. \tag{3.8}$$

A distinguishing feature of the SOM's training process is the preservation of topological properties through the use of a neighbourhood function. By updating not only the winning node, but also its neighbours, spatial relationships in the input data are recapitulated in the map via a process of global ordering (self-organisation). The SOM can therefore be seen as producing an ordered non-linear projection of data onto a lower-dimensional space, providing a generalisation of the PCA approach. The learning equation in the SOM is defined as:

$$m_i(t+1) = m_i(t) + \alpha(t) h_{ci}[x(t) - m_i(t)], \tag{3.9}$$

where t represents time, $\alpha(t)$, termed the 'learning rate', is a measure of the change effected in the model by the update ($0 < \alpha(t) < 1$), and h_{ci} defines a neighbourhood function for the winning node c. One simple choice for the neighbourhood function is based on determining $N_{cr}(t)$, a set of nodes encompassed by a radius r which is a function of time t. At any given time, if a node is within this radius ($m_i \, \epsilon \, N_{cr}(t)$), h_{ci} returns a value of 1, otherwise a value of 0 is returned. A smoother neighbourhood kernel is provided by use of the Gaussian function

$$h_{ci} = \exp\left(-\frac{|| \, r_c - r_i \, ||^2}{2\sigma^2(t)}\right), \tag{3.10}$$

where $\sigma(t)$ defines the width of the kernel at time t in a manner similar to $N_{cr}(t)$, and r_c and r_i are the locations on the lattice of vectors c and i respectively. Here, rather than a binary function of inclusion for update or not, a node's distance to the winning node dictates the magnitude of the change effected in the model, with closer nodes being subject to larger changes. The choice of learning rate and neighbourhood functions are important considerations in the design of a SOM (as discussed in [90]); both $\alpha(t)$ and the neighbourhood kernel width should decrease monotonically with time, and in order for convergence to occur it is necessary that $h_{ci} \to 0$ as $t \to \infty$. We can therefore split the training of the SOM into two broad stages based on these values; during the initial global ordering phase, $\alpha(t)$ is set to unity and the neighbourhood is quite large (usually greater than half of the lattice size), while in the convergence phase, the neighbourhood is much smaller and adjustments to the node weights are minor.

Visualisation of a trained SOM is commonly achieved by use of a U-matrix [195], a grey-scale

image where each node's value is determined by the average distance between it and its closest neighbours. Also, since the learning process may produce different maps depending on the initialisation and the sequence in which the training vectors are presented, it is useful to be able to compare the quality of learning in each case in order to determine an optimal mapping. A commonly used performance measure is the quantisation error ($|| \ x - m_c \ ||$), which indicates how similar an input vector is to its BMU. Another popular metric is the topographic error, which is calculated based on the proportion of all training vectors for which the first and second BMUs are not adjacent. This provides an indication of the 'orderedness' of the map.

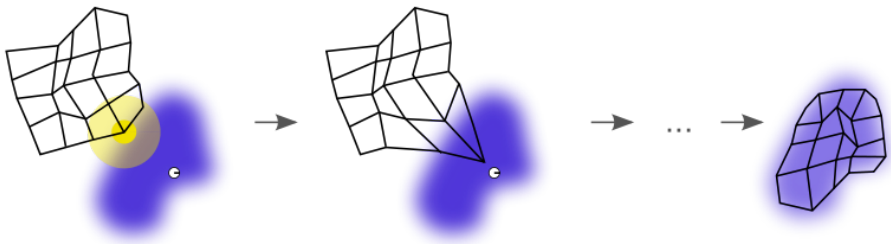

Figure 3.3: Training a Self-Organizing Map The first figure indicates the input vector in white and the BMU or winning node and its neighbourhood in yellow. The second figure shows how, after the learning phase, both the winning node and its neighbours are now closer in value to the training sample. At the end of the training process the SOM provides a good approximation of the input space with neighbouring nodes responding to vectors from similar spatial locations in the training set. (Source: Wikimedia Commons (C) Mcld. Licensed under CC-BY-SA-3.0)

The process described above is based on the original incremental learning SOM. It has been shown that a 'batch map' approach can also be used which greatly reduces computational time, as well as protects against any bias introduced by the order in which the training samples are presented to the map. In a batch learning scheme, the winning node for each vector in the competition phase is recorded, but the update procedure is not carried out until all samples have been considered. The update is then based on both the count and the mean of the subset of training samples clustered at each node j located within the neighbourhood of node i

$$m_i = \frac{\sum_j n_j h_{ji} \bar{x}_j}{\sum_j n_j h_{ji}}, \tag{3.11}$$

where h_{ji} is the neighbourhood function, n_j is the number of input vectors mapped to node j, and $\bar{x}_j$ is the mean of those vectors.

3.2.2 SOMBRERO's Approach to *de novo* Motif Finding

SOMBRERO [89, 196] recasts motif finding as a clustering problem, where short subsequences from the input set (ℓ-mers, where ℓ is the length of the motif we are searching for), are grouped by sequence similarity. Using a SOM in this way represents a completely unbiased approach to motif finding, and is an incredibly powerful technique, allowing the simultaneous characterisation of all motifs of length ℓ present in the training set. By training individual maps for multiple ℓ-mer lengths (8–32bp inclusive), it is possible to capture motifs of any length. SOMBRERO's architecture consists of an MxN lattice where the model at each node is a PWM constructed from the set of sequences clustered there. Each node also retains the PSSM ${f_{bi}}^z$ used to derive its PWM in order to simplify the learning process; the PSSM is defined as:

$${f_{bi}}^z = \frac{{c_{bi}}^z + \beta p_b}{n_z + \beta}, \tag{3.12}$$

where $z = (z_1, z_2)$ indicates the coordinates of the node on the lattice, ${c_{bi}}^z$ is a count of the occurrences of base b at position i in the ℓ-mers used to construct this PSSM, $n_z = \sum_b {c_{bi}}^z$, p_b is the background frequency of nucleotide b, and β is a scale-factor used to control the pseudocount. During training, which follows the batch map mode discussed in the previous section, the learning function allows each node's PWM to evolve in order to better portray a feature of the input space. The first step in training is to segment the training sequences into overlapping strings of length ℓ, and to assign each string x^j to its best matching node according to the log-likelihood scoring function described by equation 1.4. After clustering all of the input samples, the raw base counts at each node are updated, including the contributions from neighbouring nodes

$$\frac{\sum_{z'} \Theta(\mid z - z' \mid){c_{bi}}^{z'} + \beta p_b}{\sum_{b'} \sum_{z'} \Theta(\mid z - z' \mid){c_{bi}}^{z'} + \beta}, \tag{3.13}$$

where $\Theta(\mid z - z' \mid)$ defines a neighbourhood function which will determine the proportion of base counts that a node will contribute to another node that is a distance $\mid z - z' \mid$ away on the lattice. The neighbourhood function used by SOMBRERO is Gaussian

$$\Theta(\mid z - z' \mid) = e^{-[(z_1 - z'_1)^2 + (z_2 - z'_2)^2]/\gamma}, \tag{3.14}$$

with the γ term ($\gamma = [1/log(\delta)]$) controlling the contributions of the neighbourhood. Adjacent nodes will contribute $1/\delta$ of their counts to each other, with δ ranging from 4 to 15 during training. This ensures that initial influence from neighbouring nodes is strong but fades accordingly as training progresses. The adjusted counts are then used to generate an updated PWM and the next iteration begins. Once training has been completed (based either on a convergence criterion or a predefined

number of cycles), each substring x^j from the input sequence is assigned to its most similar node and overlaps in sequence space are resolved by only keeping the substring with the highest similarity score to the model. In order to determine which of the motifs are significant, an appropriate background model must be constructed from which statistical over-representation can be calculated. A dataset R consisting of random sequences of the same length as the input sequences is therefore generated using a Markov model of order m, with m usually equal to 3. This Markov model is based on the nucleotide frequencies in the intergenic regions of the specific organism being studied. Each substring of these sequences is then clustered at the most similar node in the trained map. This process is repeated for 100 random datasets, and a Z-score is then used to determine significance

$$z_{score} = \frac{x_{obs} - x_{exp}}{\sigma_x}, \tag{3.15}$$

where x_{obs} is the number of sequences from the input set clustered at a node, x_{exp} is the number of sequences from the random dataset clustered there, and σ_x is the standard deviation across all nodes. In order to disregard any simple repeats, a complexity filter with a threshold of 0.15 is employed

$$C(z) = (\frac{1}{4})^{\ell} \prod_{b} \left(\frac{\ell}{\sum_{i=1}^{\ell} {f_{bi}}^{z}} \right)^{\sum_{i=1}^{\ell} {f_{bi}}^{z}} . \tag{3.16}$$

The SOMBRERO algorithm is computationally costly with training time generally of the order $O(L(MN) + (MN)^2)$ for an $M \times N$ map with an input set of length L. In order to reduce this time cost, several optimisation steps were added to the implementation. When using a Gaussian neighbourhood function the contributions from distant nodes can be quite insignificant. Calculating these values is computationally wasteful; we can instead define a radius r within which the node contributions are above some threshold (such as 10^{-10} times the value of the centre) and only consider those nodes in the update function. This radius can then be decreased in line with the decrease in $\sigma(t)$ as training progresses. The second optimisation uses an updated winning node search function, in which the search for the current winning node starts with the previous winning node and its immediate neighbours. When used after an initial phase of traditional search, say, 20% of training time (by which stage the map should be somewhat ordered), this updated search function should find winning nodes more quickly. The traditional search method can then be used every tenth training cycle or so to smooth any local maxima. Finally, training set parallelisation was implemented using the Message Passing Interface (MPI) to allow the algorithm to be run on a compute cluster. Each processing node maintains a local copy of the SOM and is trained on $1/n$ of the training set, where n is the number of processors available. After each training iteration the SOM copies are synchronised. The model updates as well as the mapping of random DNA are also

parallelised.

3.2.3 Limitations and Solutions

While SOMBRERO has been shown to perform as well if not better than currently preferred motif finders such as MEME and AlignACE in both simulated and real datasets [89], there are some issues with its original implementation, which, if addressed, could improve its usefulness. The first of these issues relates to visualizing binding sites on a genome scale. Figure 3.4 displays the output of the SOMView Perl script which currently provides graphical feedback to the user after a SOMBRERO run. It displays both the trained node map (upper panel), with the nodes colour-coded by Z-score and the hits for a particular motif model in the input sequences (lower panel). The node map, while interesting from a computational point of view provides no real additional information to a biologist without further processing and also provides no indication as to the distribution and overlap in sequence space of the predicted sites associated with each of the different motif models. This approach is also clearly only amenable to a small number of short input sequences and is not suitable for genome-scale data.

Given that data from ChIP-seq experiments will include genomic locations, we modified the code to read in and keep track of these loci so that hits from all significant motif models can be then combined in sequence space. In this way, we generate peaks in a manner similar to ChIP-seq reads being mapped to a reference genome. As previously indicated, motif signals will be sampled at multiple map lengths, resulting in greater clustering depth at those loci corresponding to high confidence sites. A second pass of a peak-calling algorithm such as MACS can identify a subset of signal-rich loci, which can then be exported for visualisation in, for example, the UCSC's genome browser [126], or the Broad's IGV [127]. Alternatively, as we demonstrate in [197], a power law model (Figure 3.5) can be used to determine a cut-off of significant read cluster depth to distinguish high and low confidence sites (Figure 3.6).

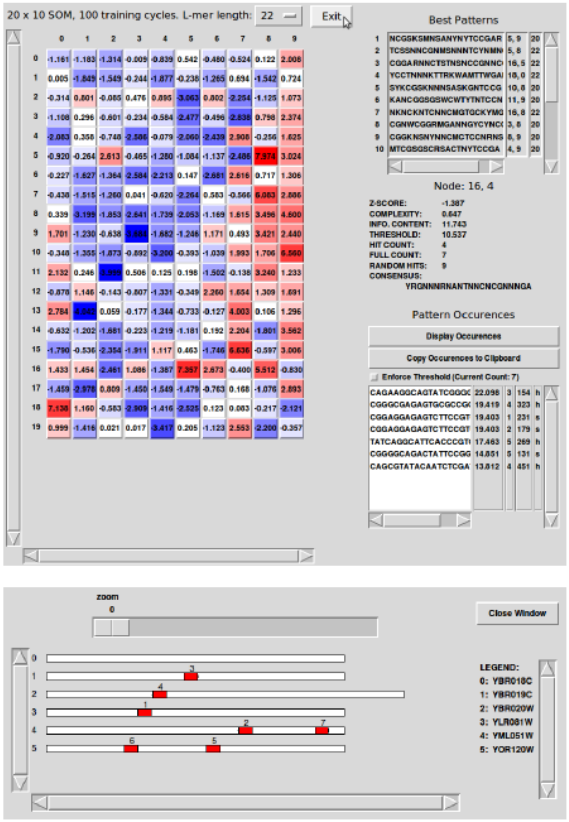

Figure 3.4: SOM Viewer A Perl script was included with the original SOMBRERO download for visualizing the trained node map and motif hits. Nodes in the map are coloured by Z-score, and clicking on a node will present the list of sequences clustered there.

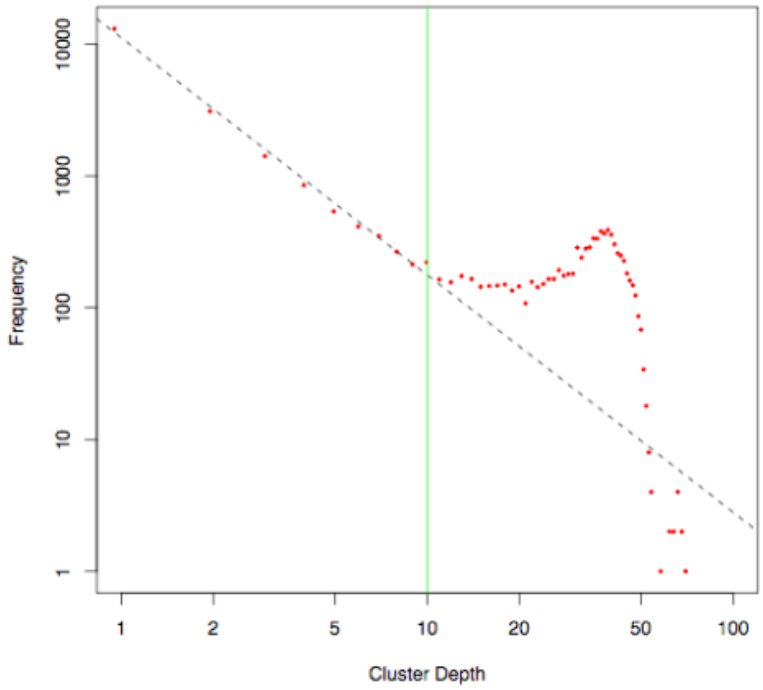

Figure 3.5: Power Law This plot shows the distribution of read frequency vs. cluster depth for all reads from the significant ChIPSOM motif models from a ChIP-chip study of CTCF binding. At lower read depths this distribution follows a standard power law; it deviaties however once a read depth of ten is reached. This read depth was therefore chosen as a cut-off to segregate peaks into low- and high-confidence subsets, as shown in Figure 3.6. (Source: This figure has previously been published in [197] and appears with permission.)

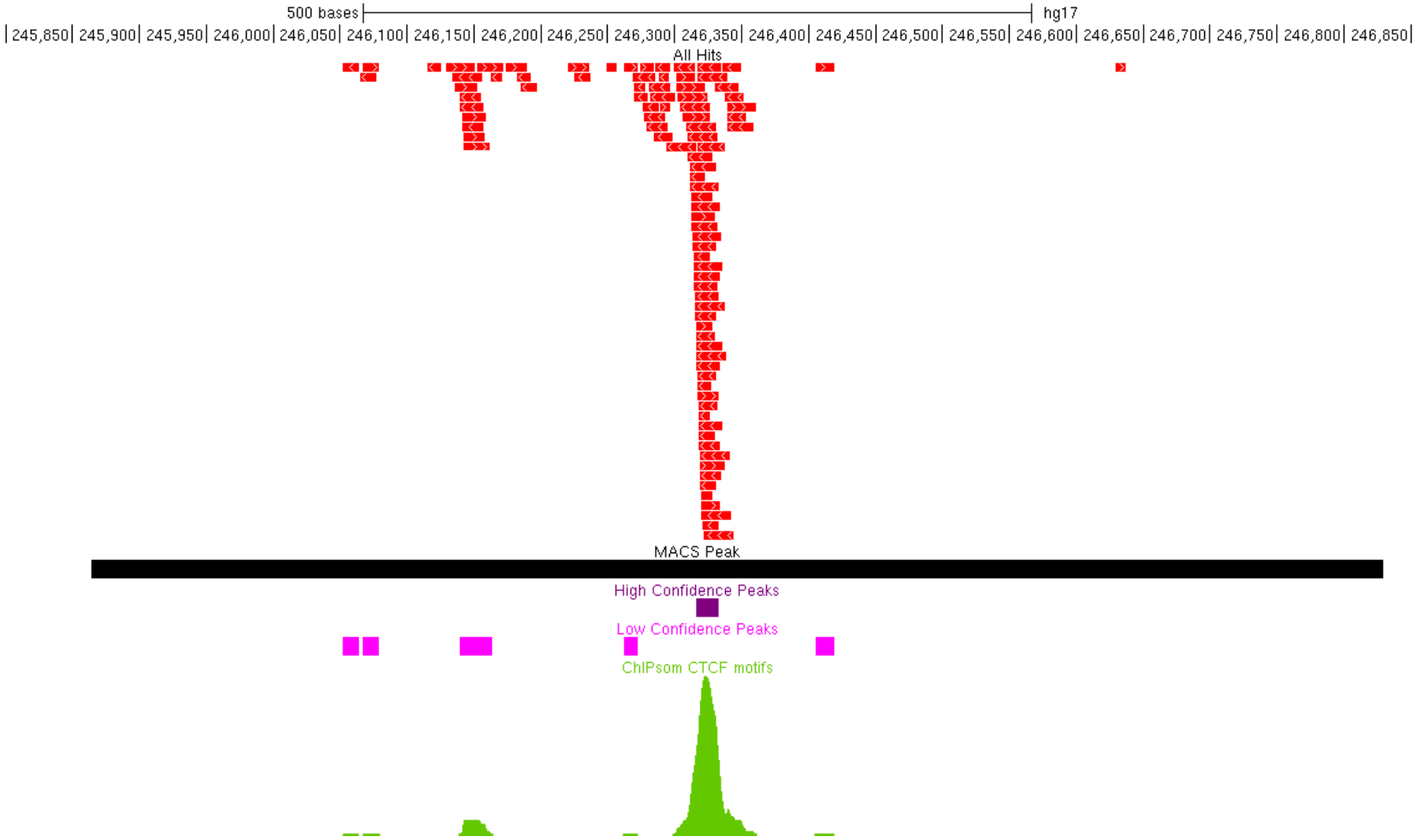

Figure 3.6: ChIPSOM Peak Segregation Here, the central black segment represents the original MACS-called peak for all mapped reads, while the red segments above it indicate the locations within this MACS peak at which sequences were matched to one of the significant ChIPSOM motif models. As can be seen, the combination of motif hits into subpeaks (green) and subsequent power law segregation into low- (pink) and high-confidence (purple) subsets can potentially result in much more specific binding site locations in the range of tens of bases as opposed to hundreds of bases.

We also sought to address the determination of a more reliable cut-off for identification of significant motifs. SOMBRERO's calculation of Z-scores provides a ranked list of nodes but leaves the user to decide on a suitable threshold for this score without any guidance. While it may be tempting to simply take any scores greater than a certain number of deviations from the mean, we believe that a more rigorous approach is to first convert the resultant Z-scores to p-values, and then to correct these p-values for the multiple testing which arises based on the fact that each individual node score represents a single statistical test of significance and the use of uncorrected values may result in a large number of Type I errors.

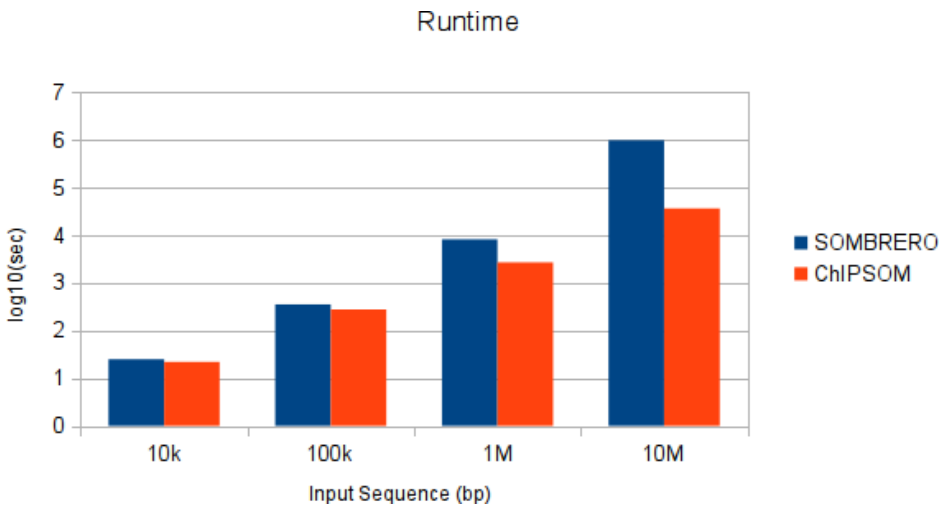

Figure 3.7: Scaling The runtime (measured in $log10$(sec)) is shown for both SOMBRERO and ChIPSOM on datasets of increasing size. While a dramatic decrease is not evident at the lower end of the input size range, for the more realistic genome-size datasets, the SOMBRERO algorithm takes ∼3 (1Mbp) to ∼27 (10Mbp) times that of ChIPSOM to run to completion. All runtimes are averages of three replicated runs using 8 processing cores of an AMD Opteron 2.2GHz cpu.

Even with the previously described steps to reduce computational cost, SOMBRERO was only ever used effectively on a limited number of promoter region sequences. There are two main reasons for this. Firstly, like all parallel algorithms, SOMBRERO is subject to Amdahl's law [198]. Generally speaking, this law describes the expected speedup in algorithm execution time as the number of processors is increased. As N, the number of processors tends to infinity, so the maximum speedup tends to $1/(1-P)$, where P is the percentage of the program that can be parallelised. In simple terms, this means that the speedup will be dictated by the serial portion of the program. SOMBRERO includes a serial step in the post-processing phase which utilises only the master node to re-scan the input sequences to ensure that all sites matching each of the final motif models are tagged as belonging to that motif. We have removed this step based on the argument that due to the use of multiple maps, even if a single motif instance is missed on any one map, *bone fide* motif instances will be shown as associated with a statistically significant PWM at more than one l-mer length and

thus will appear on multiple maps (further discussed in the next section). In order to assess the speedup in execution time gained by these modifications we carried out a series of timing experiments comparing the original SOMBRERO algorithm and the newly implemented ChIPSOM on datasets of increasing size (Figure 3.7).

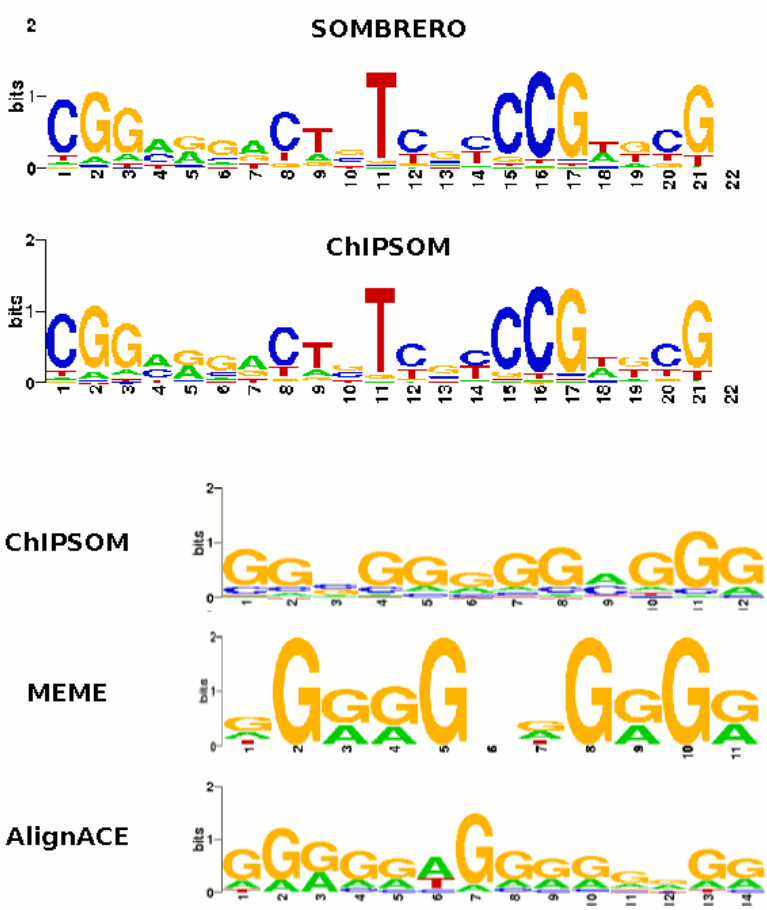

Figure 3.8: GAL4 and WT1 Sequence Logos Upper panel demonstrates that motifs returned for the GAL4 yeast promoter dataset from the original SOMBRERO algorithm and the modified ChIPSOM version are identical, supporting the elimination of the single-process functionality which results in the speedups seen in Figure 3.7. The lower panel shows the sequence logos from the WT1 ChIP-chip dataset as determined by the ChIPSOM, MEME, and AlignACE algorithms; all three motif finders produce G-rich motifs which match the canonical WT1 binding site.

While allowing us to operate on larger datasets within a more reasonable timeframe, it was essential to verify that the results of the modified algorithm were also still comparable to the original in terms of motif models returned. We therefore analyzed the GAL4 benchmark dataset which was previously used to test SOMBRERO [89] and confirmed that the returned motif models are identical

despite the changes (Figure 3.8 upper panel). We also sought to test the algorithm in terms of motifs identified when compared with results from field-leading algorithms on large datasets. Our work with collaborators in the Licht lab at Northwestern University on the genome-wide binding of the Wilms tumor 1 (WT1) transcription factor in a CCG99-11 Wilms' tumor cell line [199] provided such an opportunity. Shown in Figure 3.8 (lower panel) are the sequence logos from the application of the ChIPSOM, MEME, and AlignACE algorithms to the WT1 ChIP-chip dataset – all three motif finders returned G-rich motifs which showed strong matches to the canonical WT1 motif in both the TRANSFAC and JASPAR databases.

An important concern which we have been unable to resolve is the inefficiency of the frequentist approach used by SOMBRERO for ranking the constructed motif models. Calculating Z-scores requires generating and clustering 100 random datasets of equal size to the input dataset. If we were to consider a modest ChIP-seq dataset in which the binding protein being studied binds to several thousand loci genome-wide, with each called peak spanning several hundred base pairs, the training set size increases to a point where such an analysis becomes infeasibly time-consuming. There is also no statistical basis for this choice of 100 random datasets, providing no guarantee that true positive motif models will actually be tagged as significant. For these reasons, a probabilistic approach to determining significance may be better suited to genome-scale motif finding. It is worth noting however, that probabilistic motif finders themselves can also struggle with genome-scale datasets – MEME's runtime, for example, is cubic with respect to the number of sequences examined and quadratic with respect to dataset size in basepairs. The authors recommend that even with the parallel version (which can scale well to 128 processors), only a subset of ChIP-seq peaks (<1000) should be used for motif discovery and that expected number of occurrences should be limited to zero or one motif per sequence[2]. This restriction severely limits one of the important and more interesting aspects of secondary analysis of ChIP-seq peaks – that of finding non-canonical binding sites. One attempt to overcome this limitation (as demonstrated in Chapter Four) is to leverage General Purpose Graphic Processing Units (GP-GPUs). We have however also previously demonstrated that by first using ChIPSOM to coarsely segregate the entire dataset from a ChIP-chip or ChIP-seq experiment into subsets of high and low signal content, we can effectively pre-filter or enrich the data, allowing the subsequent use of more traditional motif finders (such as MEME) which can then operate on the entirety of the reduced dataset [197, 199].

A summary of the major differences between the previously demonstrated SOMBRERO algorithm and newly-described ChIPSOM implementation is shown in Figure 3.9.

[2]MEME user forum – http://www.nbcr.net/forum

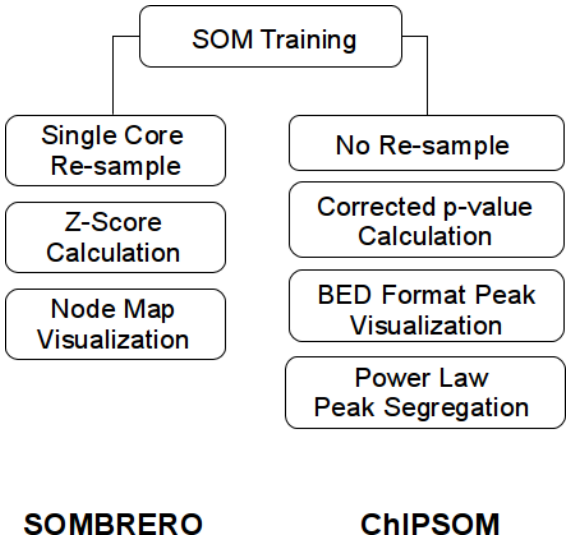

Figure 3.9: Algorithm Comparison The major differences between the SOMBRERO and ChIPSOM algorithms are highlighted. The new ChIPSOM implementation removes the re-sample step from the post-training stage. This operation (carried out on a single core) presents a significant bottleneck and its removal greatly reduces overall runtime while producing results comparable to the original implementation. The Z-score calculation step has been replaced by a more robust corrected p-value calculation making the identification of significant nodes easier. We have replaced the Perl-based node map and sequence visualisation with a more scalable BED/WIG-format peak visualisation suitable for modern genome browsers. Finally, we have included a subsequent peak segregation step which uses a power law approach to identify high-confidence peaks.

3.2.4 Motif Redundancy

As previously alluded to, a serious issue in the SOMBRERO/ChIPSOM approach to motif finding is the redundancy associated with training multiple maps of different ℓ-mer length on the same input set. By using this approach, we are essentially sampling from the same binding signal at different resolutions, resulting in a situation where motif models tagged as significant on different maps will likely represent the same core binding site with varying levels of less informative flanking sequence. This can lead to a much higher number of motif models being predicted than actually exist in the training set, reducing the usefulness of the algorithm. In order to address this problem, we propose a novel post-processing algorithm which performs a second-order clustering of the significant motif models. We present this clustering problem as an optimisation task with regard to a suitable clustering metric, and outline how a genetic algorithm (GA) approach can be used to implement this.

3.3 GMACS

In this section we look at a more general formulation of the problem of motif model clustering – that of determining familial binding profiles. We briefly discuss current state-of-the-art approaches to this problem and then introduce the concept of genetic algorithms (GAs). We provide details on the implementation of our proposed solution, GMACS, and demonstrate its performance on widely-adopted benchmark datasets.

3.3.1 Familial Binding Profiles and Current Approaches

Although SOMBRERO's inherent redundancy provided the impetus to study this motif clustering problem, more recently the concept of combining multiple motif models was discussed in [200], where the authors use this approach to determine the average binding specificity for a group of structurally related proteins. In this case, the authors manually constructed 11 profiles they term 'familial binding profiles' (FBPs) from 71 non-zinc-finger motifs taken from the JASPAR database [201]. An example FBP is shown in Figure 3.10.

FBPs are an important tool in regulatory genomics and serve a multitude of purposes: i) they can be used as informative priors for motif discovery algorithms, either biasing the search to TFs from a particular structural family, or providing a way to filter out spurious patterns and thereby increasing sensitivity [200, 202], ii) they can be used to classify novel binding proteins based on their similarity to the binding affinities of known structural families [203, 204], iii) they can be used to reduce redundancy in motif databases where minor variations or sub-motifs from the same binding site are incorrectly labelled as separate motifs; this redundancy reduction can also be applied to motif finding algorithms, either to merge similar motif predictions from a single algorithm or to combine results from multiple algorithms [205, 206], and iv) they can be used to analyze binding site turnover and provide insights into how DNA-binding mechanisms have evolved over time [207].

While [200] described the manual creation of FBPs, there have since been numerous studies which have examined metrics appropriate for motif comparison and methods for their automated clustering. Determination of motif similarity can be broadly classified into two approaches: alignment-based and alignment-free. Alignment-based methods rely on column-by-column scoring metrics such as sum squared distance (SSD), Pearson's correlation coefficient (PCC), and average Kullback-Leibler (AKL) distance. The usefulness of these metrics, amongst others, were explored in detail in [206] and expanded upon in [202]. A method for alignment-free motif comparison was provided by [208], in which the authors develop a tool (MoSta) which uses the asymptotic covariance of overlapping word sets between two motifs to determine similarity. A similar approach is taken by [209], who first convert each PSSM to a K-mer frequency vector (KFV), a 4^k-dimensional vector comprising the likelihood of each possible k-mer in a given motif, and then determine similarity using a number of distance metrics including PCC, Euclidean distance, and Cosine distance.

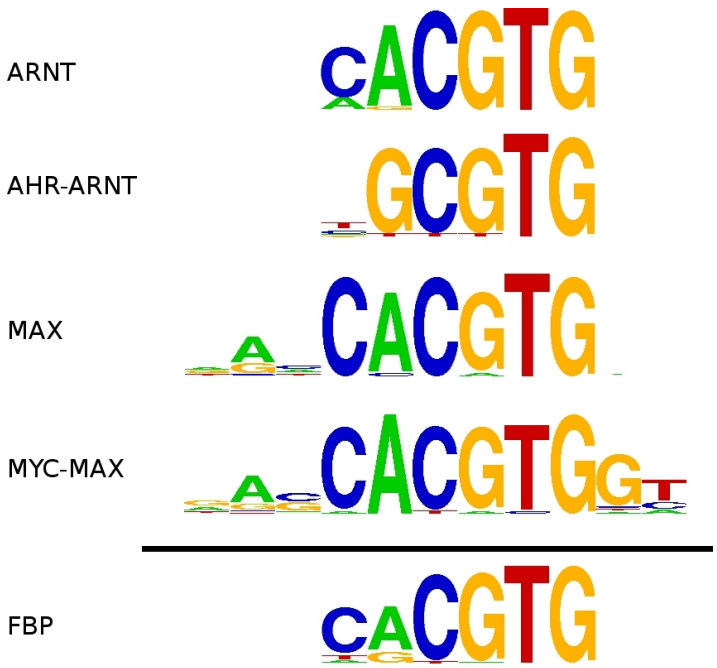

Figure 3.10: Familial Binding Profile Sample FBP for four transcription factors from the basic helix-loop-helix (bHLH) structural family. Columns which have low information content (IC) or are only present in a small number of the individual motifs are typically excluded from the FBP.

Currently, one of the most popular tools for motif clustering (as evidenced by its inclusion in the JASPAR website) is the STAMP platform [202, 210]. It offers a choice of column comparison metrics and performs pairwise gapped or ungapped local [105], or global [211] alignment, with progressive multiple alignment being performed using a UPGMA [212] guide tree. A known problem associated

this type of agglomerative approach is that it can suffer from so-called frozen subalignments [213], where a motif seemingly well-clustered early on in the tree building process is later found to better match another cluster. STAMP therefore also provides an option for iterative refinement, although this can take much longer given that each motif from the initially constructed tree must be removed and re-aligned to the remaining motifs. Ideally, we would like an approach which would allow motifs to move freely between clusters at any stage in the clustering process, thereby reducing the likelihood of convergence to a local rather than a global minimum, and obviating the need for post-clustering iterative refinement. A partitional clustering technique such as k-means [214] would provide such an approach, but is known to be highly sensitive to the effects of outliers. We therefore explore the use of the k-medoids algorithm [215] which, rather than calculating a group mean, uses one of the cluster members as the group reference. The algorithm begins with a build phase in which it selects k initial objects as the cluster medoids and then assigns each remaining non-selected object to its closest reference. In the swap phase, it interchanges selected and non-selected objects, keeping the swap if it decreases the overall sum of the dissimilarities between all objects and their corresponding cluster reference. This dissimilarity measure is termed the objective function, and the swap phase will continue until it can no longer be decreased (at which stage the best set of cluster medoids has been identified). This approach not only has the advantage of being resistant to outlier effects, but also provides the additional benefit of not having to repeatedly calculate a multiple alignment for each cluster as would a k-means approach. The k-medoids algorithm does however have two major associated problems of its own: i) like k-means, it can be sensitive to initial conditions, converging on different solutions depending on the randomly chosen starting medoids, ii) it performs a local search only, providing solutions exclusively for the given value of k; in order to fully automate the process of clustering, our approach will need to determine the optimal number of clusters for any dataset provided. To address these two issues, we propose the use of a genetic algorithm (GA).

3.3.2 Introduction to Genetic Algorithms

Genetic algorithms, based on early work by Fraser [92] and later popularised by Holland [216] and Goldberg [91, 217], are a stochastic optimisation technique making use of a population of candidate solutions. These candidate solutions, commonly encoded as binary strings (although representation as integers and floating-point numbers and are also popular), are iteratively evaluated for their effectiveness, or 'fitness', for a given function or problem domain, often termed a 'fitness landscape' or 'search space', and then combined through the use of evolutionarily inspired genetic operators such as selection, mutation, and crossover to form the next generation of candidates (Figure 3.11). A more detailed explanation of some common GA terms can be found in Table 3.1.

The parallel search capabilities of GAs (simultaneous sampling of multiple points in the fitness landscape) coupled with their ability to potentially 'escape' local minima (through the introduction of novelty via mutation) make them ideally suited to complex, noisy problem domains and their use in multiple sequence alignment [218, 219] and motif discovery [220, 221] has been well established.

GA term	Explanation
Fitness Landscape	Objective function (also known in optimisation terms as the 'problem domain' or 'search space').
Chromosome/Individual/Genotype	An array of values encoding a candidate solution to the objective function.
Gene	Subunit of a *Chromosome* encoding a particular parameter of the objective function.
Fitness	A measure of the 'goodness' of a candidate solution.
Fitness Function	Process which assigns a *Fitness* score to a candidate solution.
Generation	A single iteration of the genetic algorithm.
Population	A collection of candidate solutions during a specific *Generation*.
Evolutionary Process	Process of creating new candidate solutions from the current *Population* of solutions through the use of evolutionary operators such as *Selection* and *Crossover*.
Parents	Candidate solutions chosen for reproduction as part of the *Evolutionary Process*.
Offspring	Candidate solutions generated as part of the *Evolutionary Process* through combination of *Parent Genotypes*.
Selection	Process of choosing *Parent* candidate solutions based on their *Fitness* relative to the current *Population*.
Crossover	Process of combining *Parent* candidate solution *Genes* to generate novel *Genotypes* (*Offspring*)
Mutation	Process of adding variation to a *Genotype* which can help the GA to escape local optima.

Table 3.1: GA Terminology Explanation of some common genetic algorithm terms.

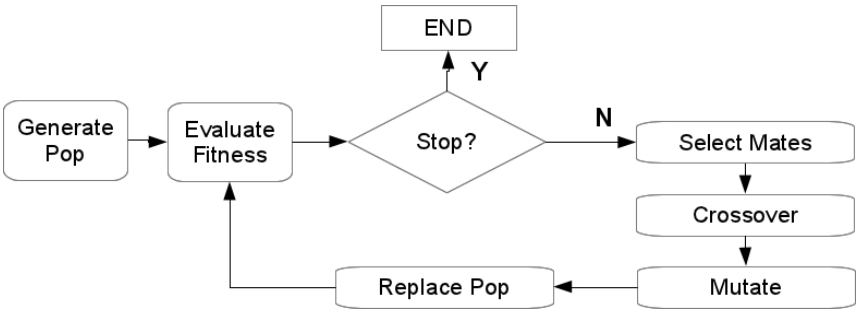

Figure 3.11: Genetic Algorithm Overview After the population is initialised, the algorithm iterates between evaluating the solutions from the current population and generating new solutions from selected candidates through various evolutionary operators. Termination usually occurs after a specified number of iterations ('generations') have elapsed, when an acceptable solution (such as within an allowable error threshold) has been found, or, after a fixed period during which the fitness has remained relatively constant.

By embedding the k-medoids algorithm within a GA framework and choosing a suitable 'fitness function' to evaluate candidates, we can leverage the local search capabilities of the k-medoids algorithm while using the GA to both perform global search for the optimal number of clusters, and to provide multiple initialisations for the k-medoids algorithm, reducing the potential impact of poorly chosen starting medoids. This type of genetic-k-medoids approach has previously been successfully applied to a number of clustering problems [222, 223].

3.3.3 GMACS Implementation

Initialisation

The first step in our algorithm is to create a matrix of pairwise distances between the motifs in our dataset. We calculate the information content (IC) for each motif position using the following equation:

$$IC = 2 + \sum_{ij} p_{ij} \cdot log_2(p_{ij}) \tag{3.17}$$

(where p_{ij} is the probability of observing nucleotide j in position i of the motif) and trim the motifs based on a user-defined threshold (default 0.3) to remove less informative edge columns. While trimming, we ensure that motifs are not shortened below a minimum length of four nucleotides. Next, we create KFVs for each of the trimmed motifs. A value of four was chosen for k based on the results reported in [209], where the authors explored various combinations of k-values and distance metrics. The choice of k=4 is also congruent with the fact that tetranucleotide frequencies have previously been shown to convey considerable genomic information [224]. We define each element,

$L_{x,m}$, in a KFV as:

$$L_{x,m} = \sum_{i=1}^{n-k+1} \prod_{j=1}^{k} (N_x)_j^T \cdot \frac{m_{i+j-1}}{|m_{i+j-1}|} \tag{3.18}$$

That is, the normalised likelihood of a k-mer, x, in a given motif, m, of length n. Here, N is a binary-encoded matrix indicating a particular nucleotide at each position in the k-mer. Once we have created our KFVs, we use the cosine distance, d_{cos}, to populate our distance matrix, where:

$$d_{cos}(a,b) = 1 - \frac{a \cdot b}{\|a\|\|b\|} = 1 - \frac{\sum_{i=1}^{n} a_i \times b_i}{\sqrt{\sum_{i=1}^{n} (a_i)^2} \times \sqrt{\sum_{i=1}^{n} (b_i)^2}} \tag{3.19}$$

and motifs with a d_{cos} close to zero are regarded as highly similar. In contrast to an agglomerative approach, these pairwise distances remain unchanged and do not need to be re-calculated as the clustering progresses and motifs are merged.

Fitness calculation

Each candidate solution in the GA population encodes K, the number of clusters in the solution, and a vector m, of length K, which holds the medoids for each of the clusters. During initialisation, the value of K for each solution is randomly chosen from the range $\{2 \dots N-1\}$, where N is the number of motifs in the input dataset. This range is chosen as the $K = 1$ solution provides no real benefit since all motifs are clustered together regardless of similarity, conversely, the $K = N$ solution places each motif in its own singleton cluster and is of little use for reducing redundancy. Once K is set, the K motifs chosen as medoids which make up vector m are also randomly chosen. Once all of the candidate solutions are initialised, we proceed to calculate the fitness for each member of the population.

We first perform one round of the k-medoids algorithm as outlined in Algorithm 1. This local search step will choose 'good' medoids based on the current value of K, and greatly speeds up the convergence of the GA towards promising solutions. Carrying out only one round of the k-medoids algorithm provides us with the benefit of improved current solutions through local search, without the computational overhead of a full k-medoids approach, which typically runs until no further updates to the medoids can be made to lower the total cost of the cluster configuration. As the goal of our GA is to both determine the optimal number of clusters and their membership, the fitness function will necessarily also include some measure of how well the data are clustered. Two methods commonly used are the Gap statistic [225] which calculates the difference between successive values of K for the test data and a bootstrapped reference dataset, and the CH-metric [226], which provides a ratio of intra- and inter-cluster distance. The authors in [207] found that a log-based equivalent

Algorithm 1 Fitness Calculation

```
1: input : pop[p]                                   ▷ member p of the current population
2: for x in 1 to motif_count do                     ▷ k-medoids
3:     assign_to_medoids(x)
4: end for
5: curr ← get_cost
6: for i in 1 to num_clusters(pop[p]) do            ▷ swap step
7:     for j in 1 to num_clusters(pop[p]) do
8:         if non_medoid(j) then
9:             swap(i, j)
10:            new ← get_cost
11:            if new < curr then
12:                update_medoids(i, j)
13:                for x in 1 to num_clusters(pop[p]) do   ▷ re-assign
14:                    assign_to_medoids(x)
15:                end for
16:            end if
17:        end if
18:    end for
19: end for
20: fit ← get_silhouette(pop(p))
21: returnfit
```

of the CH-metric (CH_{log}) was preferable to the standard metric and this log-based version was also used by [209]. Here however, we use the Silhouette metric [227], which is well-suited to partitional approaches and is defined as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3.20}$$

In this metric, $a(i)$ is the average dissimilarity of motif i to all other motifs in its own cluster, and $b(i)$, is the average dissimilarity of motif i to all motifs in its nearest neighbouring cluster. $s(i)$ is therefore an indication of whether or not an individual motif is well-placed in the clustering, or if it would be clustered more appropriately elsewhere. By creating an average $s(i)$ from each motif in the dataset, we have an overall measure of cluster 'goodness'.

Evolutionary process

Once fitness values have been assigned to each solution, they are ranked in preparation for the evolutionary process. GMACS implements a linear ranking system incorporating a selective pressure parameter, p_s, which can be used to adjust the strength of the selective bias towards fitter individuals. Linear ranking is commonly used as opposed to direct fitness values in order to avoid situations where a small number of disproportionately successful solutions leads to the premature convergence of the population. We follow an incremental or steady-state GA (SSGA) replace-worst strategy,

such that, in each generation, the bottom 5% of the parent population will be replaced by newly-created offspring. This represents a more gradual progression towards fitter solutions in contrast to a more aggressive generational strategy where the entire population is replaced at each iteration and relatively fit solutions may be lost over time due to the stochastic nature of the algorithm.

We use roulette wheel, or fitness proportionate selection, to choose the two parent solutions when generating offspring for replacement. In this form of selection (depicted in Figure 3.12), each individual in the population is assigned a 'slice' of an imaginary roulette wheel which is proportionate to its fitness within the context of the current population. The wheel is 'spun', and solutions or individuals which have larger slices of the wheel will have a greater probability of being selected for recombination. Weaker solutions will, of course, still have a small chance for selection, and this is in keeping with part of the fundamental theory of genetic algorithms, namely that part of a less fit solution's genotype may still be beneficial at a later stage when combined with genes from another solution.

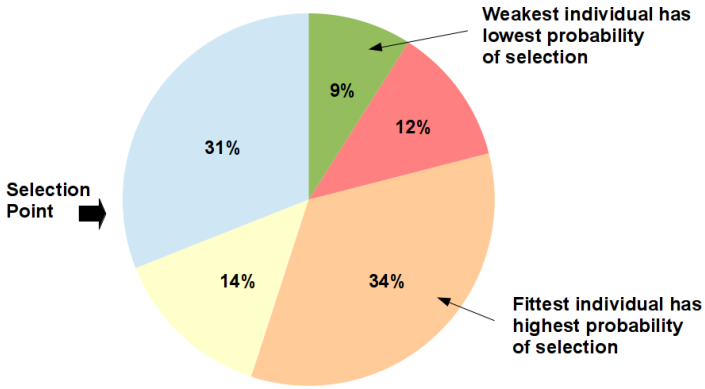

Figure 3.12: Roulette Wheel Selection Individuals are assigned a section of the roulette wheel based on their relative fitness. Here we show a sample population with only five individual solutions. Following the fitness calculation step, each of the solutions is assigned a portion of the roulette wheel corresponding to its relative fitness in the group. As larger solutions have a larger slice of the wheel, they have a higher probability of being selected to reproduce. Weaker solutions may still be selected for crossover albeit with a much lower probability.

The medoid vector representation and the effects of the crossover and mutation operators on those encodings are shown in Figure 3.13. Two parents are shown at the top of the figure, one shaded and one unshaded. Both have five clusters (a point we will return to shortly), and the index of the motif currently assigned as the medoid for each of these clusters is shown as an integer value. The form of crossover we use is termed 'uniform crossover', meaning that each separate gene in an offspring's genotype has an equal chance of coming from either parent. This type of crossover, while less common than single- or multi-point crossover, arguably produces a wider range of genotypes, exploring more of the search space. In order to explore solutions with different numbers of clusters (particularly those which may not arise as part of the random initialisation), the mutation

operator functions by perturbing the K-value for a given solution, either adding a cluster by copying the existing medoids and choosing at random an additional medoid from the remaining motifs, or removing a cluster (provided $K > 2$), by randomly choosing a medoid to delete. The probability of a mutation occurring is typically kept quite low (lest the GA risk becoming a totally random walk), and here, the rate is set at 5%.

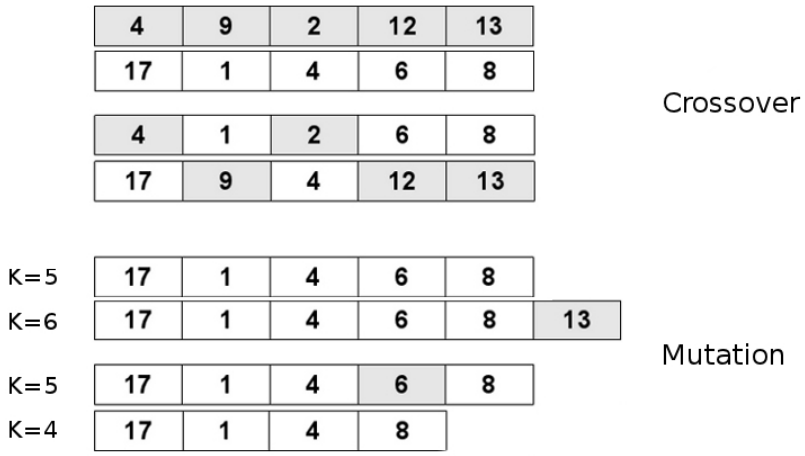

Figure 3.13: Crossover and Mutation Depiction of crossover and mutation in GMACS. The upper section shows the medoid vector representation of two selected parents, one wholly-shaded and one wholly-unshaded. Shown below them are the two offspring resulting from their uniform crossover. The lower section demonstrates the two modes of K-value mutation: addition or removal of a randomly selected cluster, shown as shaded.

As individuals in the population may have different K-values, special consideration must be given when carrying out the selection step. During the k-medoids phase of the fitness calculation, the current set of medoids is updated to a partially-optimised state. Crossover of medoids between solutions containing different numbers of clusters would result in a disruption to this improvement. If the algorithm were carrying out a full k-medoids implementation this would not present any problem since the medoids would be optimised on the next pass of the fitness function, since however, only one pass through the medoids occur, crossover is constrained to individual sharing the same number of clusters. Figure 3.14 shows the modified selection process to account for this fact. Once the first parent is selected, a check is made to see if there are any other individuals in the population with the same number of clusters – if there are, then the second mate is selected from within that subpopulation and crossover occurs as normal. If the individual, however, is the only member of the population with that specific value of K, then no valid mate exists and the crossover step is skipped. This figure also show two additional features of the algorithm design. The first of these is

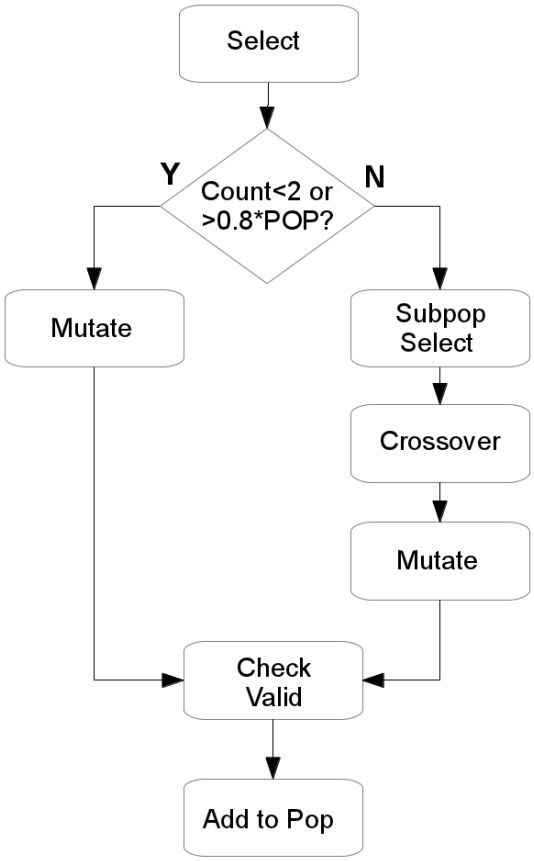

Figure 3.14: Modified Selection Process This modified selection process is designed to only allow the recombination of individuals sharing the same K-value. In cases where no suitable mate exists, or where the pre-determined diversity threshold has been exceed, mutation will occur (with the standard probability) without crossover.

the concept of population diversity, expressed as the proportion of the population with the same K-value. When this value is greater than a pre-defined threshold (default: 0.8), it gives an indication that the population has largely converged on a solution with a specific number of clusters and mutation is increased to maintain diversity and encourage further exploration of cluster space. The second feature is the offspring validity check which is necessary after crossover and/or mutation to ensure that there are no duplicate medoids as a result of the recombination or mutation.

3.3.4 Results

Motif Comparison Methods

Our first dataset consists of 355 motifs from the six largest structural families in the TRANSFAC [228] database and has previously been used by [208, 229, 209], and [207] to benchmark retrieval

accuracy, or the ability of different metrics to distinguish between motifs of different structural classes. The accuracy is based on the number of times the closest match returned from the database was of the same structural class as the query motif. GMACS and the KFV approach achieves the highest average retrieval accuracy of 0.90, compared to 0.87 for the word covariance approach of MoSta, 0.87 for the STAMP platform when using PCC and ungapped local alignment, and 0.86 for Bayesian approach outlined by the authors in [229] in which they construct a multi-class classifier with feature selection by applying sparse multinomial logistic regression (SMLR) to feature vectors of length 1390 incorporating measures such as various nucleotide frequencies, presence of palindromic features, and previously published submotifs (Table 3.2).

	GMACS	STAMP	MoSta	Bayesian
bZIP (93)	0.92	**0.94**	0.90 (**0.94**)	0.92
C2H2 (74)	**0.82**	0.76	0.76 (0.72)	0.77
C4 (52)	**0.98**	**0.98**	**0.98** (0.94)	0.91
Homeo (50)	0.88	0.82	0.82 (**0.92**)	0.85
Forkhead (49)	**0.92**	0.90	**0.92** (0.86)	0.83
bHLH (37)	0.89	0.81	**0.92** (0.73)	0.88
Total (355)	**0.90**	0.87	0.88 (0.86)	0.86

Table 3.2: Retrieval Accuracy The ability of the KFV metric to distinguish different structural classes of motif is compared to three alternative approaches using the TRANSFAC benchmark dataset. GMACS scores the overall highest average accuracy and does particularly well on the complex C2H2 zinc-finger family which causes problems for some of the other approaches. MoSta scores are Smax (Ssum) which represent the log-odds ratio of the independent and overlap probabilities of hits from two motifs on a given DNA sequence.

Genetic-k-Medoids Clustering

We demonstrate the clustering performance of our algorithm on 79 motifs from the JASPAR database. This dataset comprises the 71 motifs used by [200] in their initial manual creation of FBPs, plus a further eight zinc-finger proteins, four from the DOF zinc finger protein (DOF) family, and four from the GATA binding protein (GATA) family. We compare our results to those reported by the authors of STAMP [207] and MoSta [208] who use the same dataset to benchmark their approaches. We report the results both in terms of number of clusters defined, and structural homogeneity of the created clusters. The ability to create structurally homogeneous clusters is key to the generation of biologically meaningful FBPs, and as we show in the next section, GMACS performs very well in this regard, successfully segregating even distinct subtypes within certain structural families. For STAMP, PCC was used with ungapped local alignment and a UPGMA guide tree (default settings). The authors of MoSta provide their own clustering approach whereby they select and merge motifs based on their word covariance similarity metric. GMACS settings for this experiment were: population size equal to 100 and number of generations for which to evolve the population

equal to 300. The default mutation rate of 0.05 and information content threshold of 0.3 were used – this trim threshold was consistent with the same parameter setting in the STAMP algorithm.

	GMACS	STAMP	MoSta
Homogeneous	13	9	11
Heterogeneous	5	7	3
Singletons	0	2	12
Total	18	18	26

Table 3.3: Cluster Summary Number and type of clusters automatically determined by GMACS, STAMP and MoSta for the 79 motif JASPAR dataset originally grouped into 11 FBPs manually by Sandelin and Wasserman.

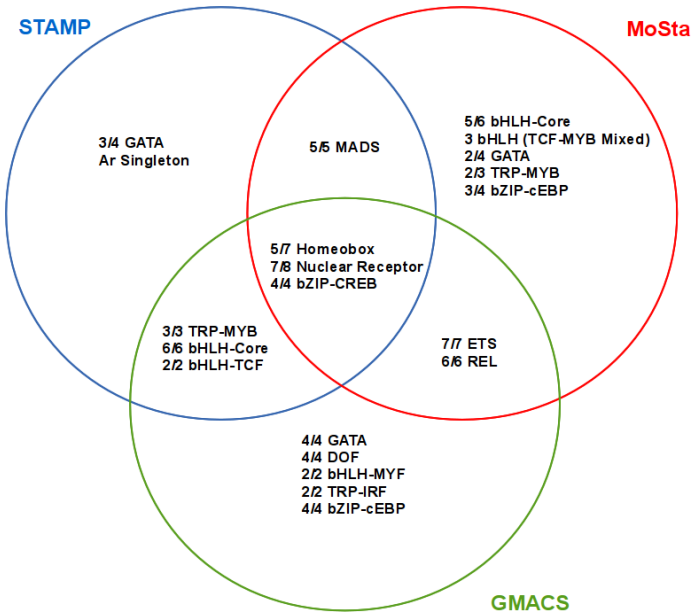

Figure 3.15: Cluster Overlap A Venn diagram displays the overlap in terms of shared homogeneous clusters between the three motif clustering algorithms.

We first provide an overall view of the solutions provided by each algorithm before examining some of the differences in detail. In total, MoSta produces eleven homogeneous and three mixed clusters, containing 67 of the 79 motifs. STAMP estimates the number of clusters at eighteen, producing nine homogeneous, seven mixed and two singleton clusters (Table 3.3). GMACS also produces eighteen clusters, but thirteen of these are homogeneous, while the remaining five contain motifs from different structural families. The twelve singletons produced by MoSta can be attributed to the inclusion of a similarity threshold in their clustering approach which prevents the merging of motifs if an FBP becomes too heterogeneous as a result of the merge. The homogeneous clusters

shared by each of the algorithms are depicted in Figure 3.15.

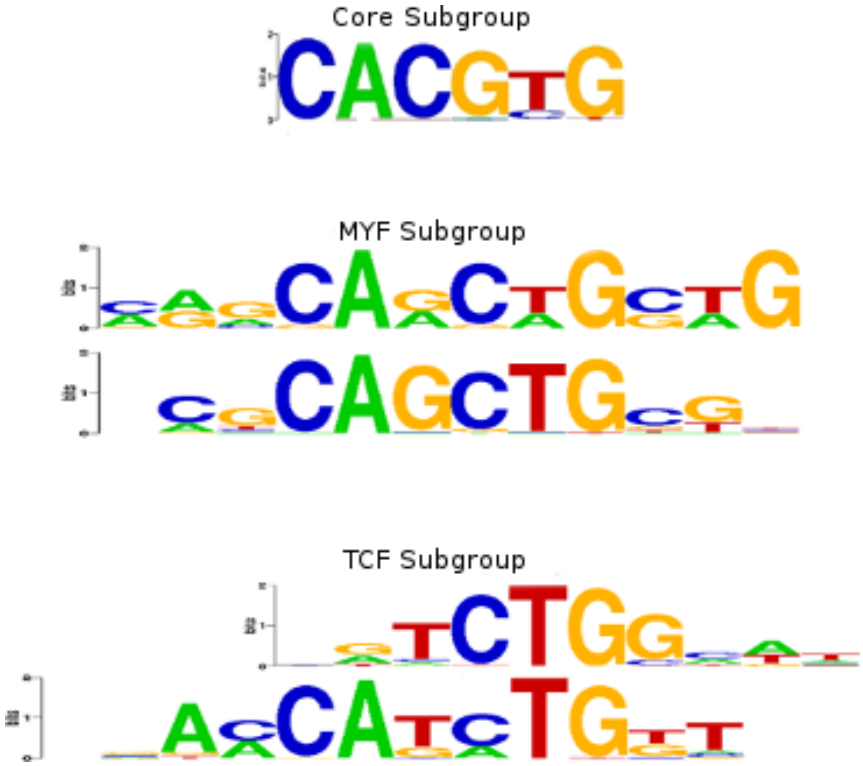

Figure 3.16: bHLH Motif Family The bHLH familty of motifs from the JAPAR dataset comprises three distinct binding subtypes. GMACS correctly classifies the motifs into these three subtypes, the core subtype containing six motifs (ARNT, AHR-ARNT, MAX, MYC-MAX, USF, n-MYC), and MYF (MYF, NHLH1) and TCF (HAND1, TAL1) subtypes, each containing two motifs.

We begin our detailed examination of the results with the ten members of the bHLH family which form three distinct subgroups (as shown in Figure 3.16). STAMP creates two homogeneous clusters with six and two members respectively. Of the remaining members, one is clustered with the GATA1 zinc-finger (GATA1) and forkhead box L1 (FOXL1) motifs, while the other is clustered with the E26 transformation-specific (ETS) family motifs. MoSta groups the motifs as a cluster of five and three. The larger cluster is, rather surprisingly, missing the aryl hydrocarbon receptor-aryl hydrocarbon receptor nuclear translocator (AHR-ARNT) motif which contains the strong consensus 'CACGTG' sequence associated with that subgroup, while the smaller cluster includes the T-cell acute lymphocytic leukemia 1-transcription factor 3 (TAL1-TCF3), nescient helix-loop-helix 1 (NHLH1), and myogenic factor (MYF) motifs, mixing the remaining subtypes. GMACS provides the only approach to create three homogeneous clusters. The first cluster is the same six-member group created by STAMP, the second contains the NHLH1 and MYF motifs (MYF subgroup), while the final cluster groups the TCF subgroup motifs TAL1 and heart and neural crest derivatives expressed 1 (HAND1)

together. All three of the algorithms separate the four bZIP cAMP response element-binding protein (CREB) subgroup motifs into a homogeneous group as well as clustering seven of the eight nuclear receptor motifs together. The remaining androgen receptor (AR) motif is classed as one of two singletons by STAMP while GMACS clusters this motif with the two homeobox motifs, engrailed homeobox 1 (EN-1) and paired box 4 (PAX4). It is possible that the length and complexity of the AR and PAX4 motifs play a role in this particular grouping, affecting the number of shared k-mers between the two.

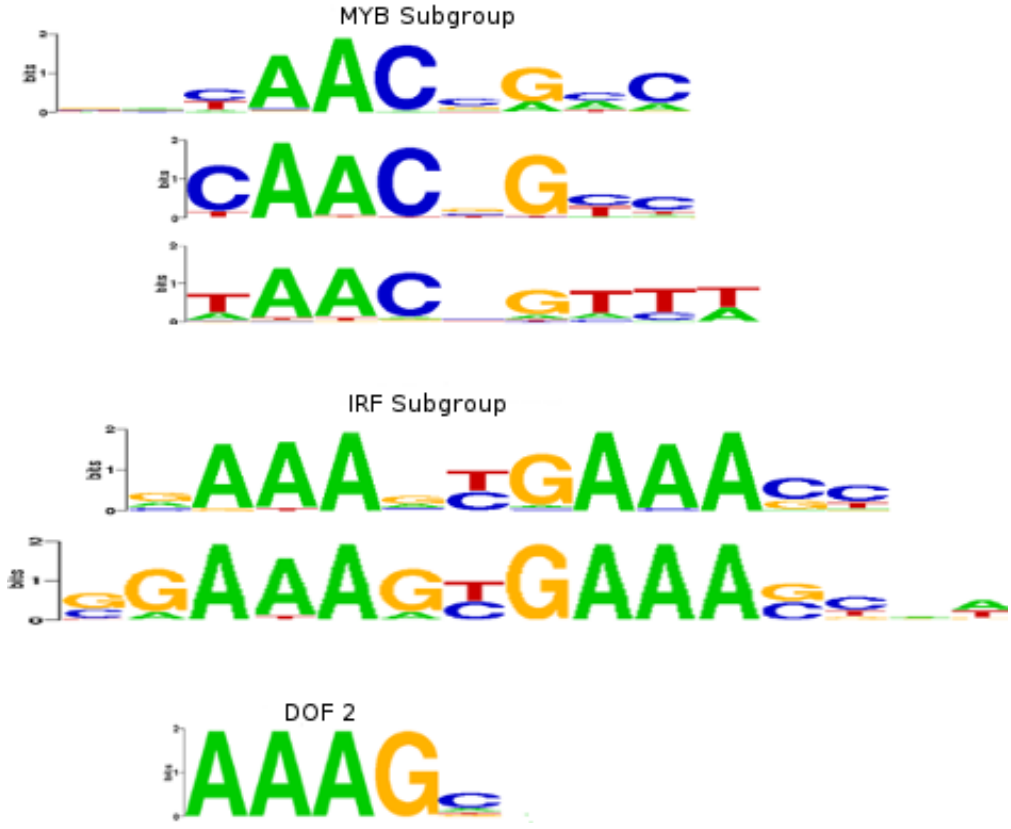

Figure 3.17: TRP Motif Family The TRP family of motifs comprises two binding subtypes. The first of these is the MYB group which includes three motifs (GAMYB, c-MYB, and MYB.PH3), while the second subgroup is made up of IRF1 and IRF2 (trimmed here for display purposes). Both STAMP and MoSta include a DOF family motif (an example of which is shown here) with the IRF subgroup based on a strong 'AAAG' signal.

The TRP group of motifs contains two distinct subfamilies. The first of these, the v-myb avian myeloblastosis viral oncogene homolog (MYB) group, are recognised by STAMP and GMACS as a homogeneous cluster of three motifs, GAMYB, a gibberellin- and abscisic acid-regulated MYB

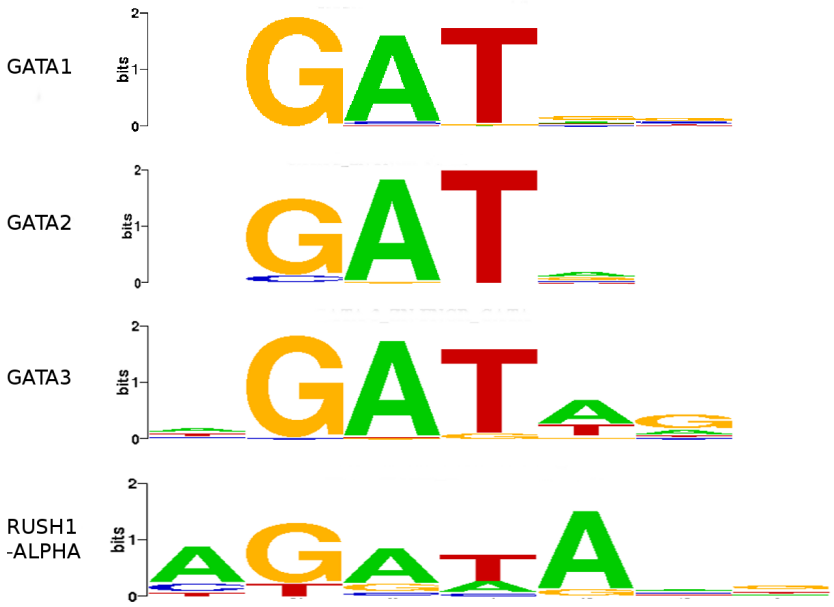

Figure 3.18: GATA Motif Family The four GATA motifs include the canonical 'GATA' consensus and are group together by GMACS. STAMP mis-clusters one of these with the TAL1 and FOXL1 motifs, while MoSta classifies two of the four as singletons.

found in maize, c-MYB, and MYB.Ph3 (found in Petunia hybrida). MoSta, on the other hand, only recognises two of these motifs as belonging together, excluding MYB.Ph3 from this cluster. The second TRP subfamily is comprised of the interferon regulatory factor 1 and 2 (IRF1 and IRF2) motifs. While both STAMP and MoSta group these two motifs with the four DOF zinc-finger motifs as a single heterogeneous cluster, GMACS instead creates two homogeneous clusters. The clustering by STAMP and MoSta in this case is reasonable however, given the strong 'AAAG' DOF family motif signal which is easily mistaken for a submotif of the IRF family (Figure 3.17). Both MoSta and GMACS cluster all of the ETS and v-rel avian reticuloendotheliosis viral oncogene homolog (REL) family motifs separately as homogeneous groups. The STAMP ETS cluster however, also includes the HAND1-TCF3 bHLH motif making this group heterogeneous, while its REL group contains the bZIP CCAAT-enhancer-binding protein (cEBP) subgroup motif cEBP homologous protein (CHOP-cEBP). This cEBP subgroup is split into two clusters by GMACS, one is homogeneous and contains the cEBP and CHOP-cEBP motifs, while the other contains the nuclear factor, interleukin 3 regulated (NFIL3) and hepatic leukemia factor (HLF) cEBP motifs as well as the forkhead motif FOXC1. GMACS also incorrectly clusters a single forkhead motif, FOXL1 with the five members of the MADS-box family whereas STAMP and MoSta maintain the MADS group as a homogeneous cluster. This inclusion of FOXL1 with the MADS-box family may be explained by the shared 'TATTTAT' sequence.

The clustering of the four highly-conserved zinc-finger GATA family motifs (Figure 3.18) shows

considerable variation among the three algorithms. MoSta does poorly, clustering only two of the four motifs together. STAMP clusters three of the four while the remaining member is, as previously indicated, clustered with TAL1 and FOXL1. GMACS however, creates a homogeneous cluster from the four motifs. Our final set of motifs includes members from the homeobox, high-mobility group box (HMG) and forkhead families. Firstly, all three approaches cluster five of the seven homeobox motifs into one homogeneous cluster. STAMP clusters PBX with the four Sry-related HMG box motifs, SOX17, SOX19, SOX5 and SRY, and creates a single combined HMG/homeobox/forkhead cluster comprising both these motifs and six motifs from the forkhead family. STAMP also creates a HMG/forkhead group containing HMG-1 and FOXC1, and a HMG/homeobox group containing HMG-IY and PAX4. MoSta clusters the four HMG motifs with the six forkhead motifs as in the case of STAMP, but does not include the PBX Homeo motif. It also creates a HMG/homeobox group but in this case containing HMG-1 and EN-1. GMACS, like STAMP, clusters the PBX homeobox motif with the four HMG motifs, but as a separate cluster from the six forkhead motifs, which are instead clustered with another set of HMG motifs: HMG-IY and HMG-1.

FBP Construction and Stability

Once we have clustered the motifs, we must generate the FBP for each of the defined clusters. This can be achieved through any of the standard multiple alignment methods, although it has previously been shown that a local Smith-Waterman alignment may be preferred for binding motifs which are typically short ungapped sequences [207]. The membership and FBPs for each of the clusters derived by GMACS for the JASPAR dataset are shown in Figure 3.19. In order to assess the stability of the FBPs in our solution, we perform a leave-one-out cross-validation (LOOCV) as carried out in [207, 208]. Each motif in turn is removed from its cluster, the cluster FBP is re-calculated minus the contribution of the test motif, and then the motif is re-aligned to each of the FBPs. If it is re-assigned to its original cluster, we consider the classification a success. The LOOCV rate for the 79 JASPAR motifs is 96% for GMACS, compared to 91% for STAMP. The improved classification rate for our approach is unsurprising given that more of the clusters elucidated are structurally homogeneous and therefore FBPs are less likely to be affected by the removal of any individual motif. The authors in [208] successfully manage to re-cluster all of their 67 clustered motifs to their original FBPs, which, again, is unsurprising given the high number of singletons in their solution.

3.4 Discussion

ChIPSOM and GMACS

In this chapter we have presented ChIPSOM and GMACS, tools for the automated discovery and clustering of binding motifs in high-throughput ChIP data. ChIPSOM addresses many of the limitations of the original SOMBRERO algorithm, allowing it to scale to larger, genome-scale datasets.

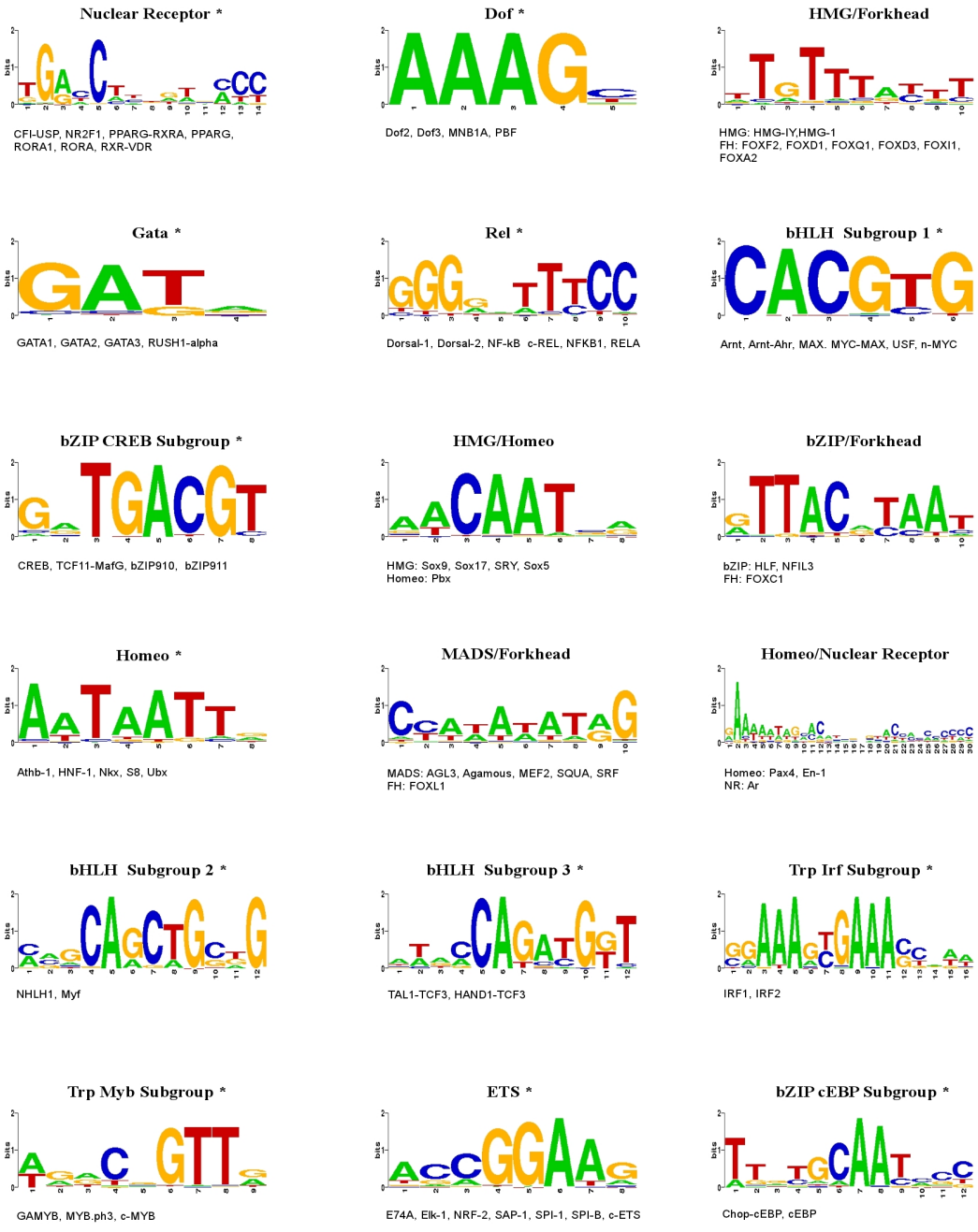

Figure 3.19: Final Cluster Composition Cluster membership and FBPs for the eighteen clusters identified by our method in the test dataset of 79 JASPAR motifs. Clusters marked with an asterisk are structurally homogeneous.

We have demonstrated its improved visualisation capabilities and compute time, as well as its application to ChIP-chip datasets. In Chapter Five we will demonstrate its application to data from a ChIP-seq experiment.

GMACS was designed to address the problem of motif redundancy inherent in the use of multiple maps during motif discovery. It also attempts to circumvent the problem of frozen sub-alignments associated with tree-based FBP techniques. Our results demonstrate the K-mer Frequency Vector to be a suitable metric for comparing motifs in an effort to establish common binding affinities based on structural family, and have also provided evidence that a genetic-k-medoids approach may be used to both determine cluster number and membership. We must however recognise some weaknesses arising from our approach. A common concern with genetic algorithms is that they can be computationally costly, with the most compute-intensive task usually being the evaluation of the fitness function. The k-medoids algorithm is also computationally intensive given that the swap stage typically progresses until no further exchanges can be made to decrease the overall cost of the cluster configuration. While the complexity of the standard algorithm is $O(k(n-k)^2)$ (where k is the number of medoids and n is the number of objects to be clustered), we have shown that a single round of local search using the k-medoids as part of the fitness evaluation function is enough to greatly reduce the number of generations necessary for the GA to converge on good solutions. The silhouette component of the fitness function however requires an all-to all comparison adding a $O(n^2)$ term to GMACS' overall complexity. This indicates that while increases in other parameters such as number of generations and population size will invariably increase the runtime, the rate-limiting step will inevitably be the number of motifs in the input dataset. For the relatively small test dataset of 79 JASPAR motifs, and setting a population size of 100 and generations for which to evolve equal to 300, the average time to completion on a 2.6 GHz Intel Xeon processor calculated over 100 runs was 7.97 seconds. Future work will aim to improve on this through further optimisations of the code and parallelisation using OpenMP.

While GMACS aims to convergence to a global rather than a local minimum, like all GAs, it is a stochastic algorithm and there are therefore no guarantees that it will in fact achieve this goal. In order to test its convergence properties, we ran the test dataset of 79 JASPAR motifs 10,000 times to ascertain the number of times which the algorithm would converge on any particular solution. The K=18 solution (fitness: 0.469) which we have presented here accounts for $\sim$90.5% of the solutions returned. The second most common solution, accounting for a further 5% of the explored cluster configurations, is a K=19 (fitness: 0.465) solution in which the FOXC1 motif is classified as a singleton, resulting in the NFIL3 and HLF cEBP motifs becoming a homogeneous cluster. The third most common solution occurs in $\sim$1.2% of the runs and is another K=19 solution, also involving a FOXC1 singleton. This time however, the forkhead group becomes homogeneous and the two HMG motifs from the previous HMG/forkhead group are re-clustered elsewhere. The fitness for this third solution is in fact slightly higher (0.471) than that of the K=18 solution, raising several important points. Firstly, it illustrates the fact that the GA will quickly converge on good but not necessarily optimal solutions. This result also points to the difficulty, not only for GAs but also for most algorithms operating in complex problem domains, of appropriate parameter selection.

The trim threshold, mutation rate, population size and number of generations, for example, will all play a role in the type of solutions returned. One promising approach to addressing this issue, and deserving of further examination, is the use of adaptive GAs, where the parameters themselves can evolve as the algorithm progresses. Finally, the higher fitness of the K=19 solution may also be an indication that a modified or alternative fitness function and/or set of evolutionary operators could help us to move the GA towards these less commonly explored regions of the solution space.

While being cognisant of these issues raised, we still posit that our algorithm is a useful and effective alternative to current standard approaches. The most common clustering solution provided by GMACS for the benchmark dataset is both consistent and biologically meaningful, comprising a larger number of structurally homogeneous clusters than either STAMP or MoSta, without the need for a large number of singleton clusters in order to achieve this.

SOM Variants and Related Neural Networks

The batch map used in SOMBRERO is just one of a multitude of variants on the original SOM algorithm which exist. Here, we consider some of these variants, as well as discuss potentially useful alternative neural network architectures which address some fundamental issues in the SOM algorithm.

The hierarchical, or multilayer SOM (e.g. [230],[231]), is an extension to the architecture which uses successive layers of SOMs in a pyramid shape. The first, or top-layer SOM is trained on the entire input set, while SOMs at subsequent layers are trained on the subset of samples clustered at a single node (or within a cluster of nodes) from the layer above. This progressive filtering of vectors both allows a user to explore the data in a multilayered way (showing global ordering within the data as well as subtle relationships among identified subclasses), and reduces the computational cost associated with training the map.

A common concern when training SOMs is the issue of choosing a suitable map size. If a map is too small, overcrowding at the nodes can occur, resulting in a loss of performance in terms of cluster separation. Conversely, starting with a map larger than needed, or running multiple maps to ascertain the optimal mapping can be computationally wasteful. One approach which seeks to address this problem involves the use of growing SOMs. This is a class of SOM algorithm which supports the dynamic expansion (or contraction) of the lattice during training, based on perceived need. Two main types of growing SOM exist – those that grow on a fixed grid and those that grow freely. The Growing Grid [232], for example, maintains a rectangular topology and grows in the following manner. First, the most active node on the lattice q is determined – this is the node that has been identified as the BMU most often. Each neighbouring node of q is then examined to determine which of them has the most divergent model from it, the selected node is labelled f. Finally, either a row or column of nodes (depending on the location of f) is inserted between q and f; the reference vectors for these new nodes are interpolated from their neighbours. The underlying assumption here is that f indicates a direction with high variance in the input data, and adding the extra nodes will better distribute the signals resulting in an improved separation of the distinct

clusters. An important subclass of constrained growing SOMs for bioinformatics are algorithms such as the Competitive Neural Net [233], self-organizing tree [234], and self-organizing tree algorithm (SOTA) [235]. These algorithms provide a way of limiting a SOM's growth to a binary tree format and therefore have applications in evolutionary analysis and phylogenetic reconstruction. Examples of unconstrained growing SOMs include the growing cell structure (GCS) [236] and growing neural gas (GNG) [237, 238]. The GCS uses k-dimensional hypertetrahedons as basic building blocks for the network – lines for $k = 1$, triangles for $k = 2$ etc. The vertices of the tetrahedrons are the neurons and the edges represent neighbourhood relations. Key differences from Kohonen's algorithm include the fact the neighbourhood width around the BMU is kept constant, and only the BMU and its direct neighbours are updated. Node insertion is carried out after a predefined number of adaption steps and is based on a concept of 'resource'. In a manner similar to the Growing Grid, the node which receives the highest number of input vectors is selected and a new node is inserted between it and its neighbour with the most dissimilar model. Nodes which do not receive many hits can be deleted during the growth phase, sometimes resulting in the creation of several smaller subnets which continue to grow and split independently. Each of these subnets represents a distinct cluster in the input space, and Fritzke has demonstrated [239] that this can lead to a better characterisation of the subtleties of the underlying input data. While disjointed and dynamic-topology SOMs have been described, Kohonen points out that the original SOM was "conceived for non-parametric regression whereupon it was considered more important to find the main dimensions ... in the signal space".

There are also many algorithms that, while not directly extending the SOM, are highly related. Learning vector quantization (LVQ) is one such algorithm (also proposed by Kohonen [90]), which can be used to perform supervised vector quantisation or classification. With LVQ, class labels are associated with each input vector x, and the learning rule is updated to take advantage of this information as follows:

$$m_i(t+1) = m_i(t) + \alpha(t)s(t)\delta_{ci}[x(t) - m_i(t)]. \tag{3.21}$$

Here, $s(t)$ has a value of $+1$ if x and m_c belong to the same class (form part of the same Voronoi set) and a value of -1 otherwise. The Kronecker delta δ_{ci} has a value of 1 if $c = i$ and a value of 0 otherwise. The key differences from the SOM algorithm are therefore i) only the BMU is updated, there is no update of neighbouring nodes, ii) model, or codebook vectors are only moved closer to the clustered input vectors if they share the same class label (i.e. the training inputs are correctly classified), and iii) codebook vectors are made less similar to clustered input vectors if they do not share the same class label (i.e. the inputs are misclassified).

A problem associated with competitive learning is the notion that patterns learned early on in the training process may not persist if similar data are not seen at a later stage. The question then becomes how to implement a system which can preserve previously learned knowledge yet still retain the ability to learn new patterns – this is known as the stability-plasticity dilemma, and was posed by Carpenter and Grossberg who went on to detail a neural network solution to this problem they

call adaptive resonance theory (ART) [240]. Although many variations of the ART algorithm have been produced [241, 242, 243], here, we will focus on the initial binary design, ART1.

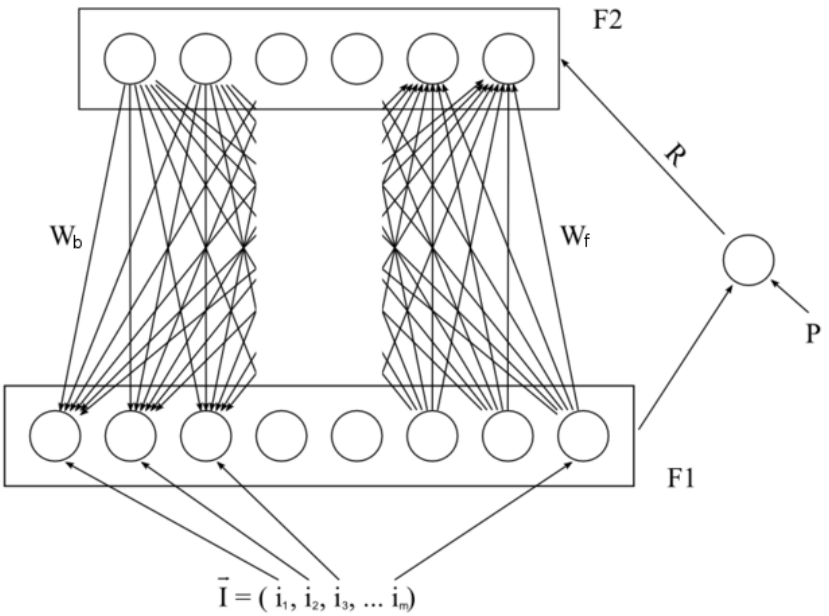

Figure 3.20: Adaptive Resonance Theory (ART1) A simplified architecture for the ART1 neural network depicting the F1 and F2 layers, forward (W_f) and backward (W_b) weights, and reset module R, with vigilance parameter P. Gain controls not shown. (Source: Wikimedia Commons (C) Christoph Mller. Adapted and licensed under CC-BY-SA-3.0)

An ART1 network (Figure 3.20) consists of two layers of neurons – F1, the comparison layer, and F2, the recognition layer, containing the class templates, or previously-learned categories. The network also incorporates two gain controls, G1 and G2, and a Reset module. All neurons in the F1 and F2 layers are connected by weights which multiply the signals in the continuous-valued forward (W_f) and binary-valued backward (W_b) long term memory (LTM) functions. In the recognition, or bottom-up phase, the neurons in the F2 layer integrate the signals from the F1 layer in a winner-takes-all competition, with the winning neuron inhibiting signals from the other nodes through lateral inhibition. In the comparison, or top-down phase, every node in the F1 layer receives three input signals: the input vector, a feedback signal (W_b) from the winning F2 neuron containing the matching class template, and a gain (G1) signal. Only those neurons which have high input signals

from two of the three sources are activated (the 'two-thirds rule'). A vigilance test is then applied using the inner product of the input vector and the feedback signal from F2. If the match is above a certain threshold (specified by the vigilance parameter P), then a previously-learned category has been recognised and learning occurs with the weights being updated accordingly. If a mismatch occurs, the reset mechanism is used to inhibit the signal from the winning F2 neuron (mismatched category) and the next best-matching neuron is tested. This process is repeated until either the vigilance test is passed, or all F2 nodes are exhausted. If no matching class template has been found, then a new node is initialised with the input pattern serving as the new template. It can be seen that the vigilance parameter controls the granularity of clustering – a low threshold will result in a small number of large clusters with a higher likelihood of mis-classifications, while a high threshold will result in a larger number of small specific clusters.

Generative topographic mapping (GTM) was introduced by Bishop et al. [244] in response to several criticisms of the SOM algorithm. These include: i) SOM parameters such as learning rate, lattice size, and neighbourhood function being largely heuristic and having no theoretical basis ii) the SOM not implementing a global optimisation function, nor explicitly defining a probability density function, and, finally, iii) convergence of the weight vectors not being guaranteed. The GTM therefore provides a probabilistic approach, using EM to learn the parameters of a latent variable model, where the latent space is defined as a discrete grid of points (similar to the SOM lattice) which is assumed to be projected in a non-linear manner into the input space. By assuming Gaussian noise in the input space, the model becomes a constrained mixture of Gaussians. This approach has the benefit of an explicitly formulated density model over the data, where convergence is guaranteed using a well defined probabilistic optimisation technique, and is measurable by an associated cost function.

It was first argued by Linsker that maximizing the average mutual information represents an optimal way to extract statistically salient features from an input signal [245]. This idea forms the basis of work by van Hulle to develop the maximum entropy learning rule, which, when applied to an input signal, results in the creation of an equiprobabilistic topographic map which can be used for density estimation [246]. This work was further developed with the introduction of the kernal-based maximum entropy learning rule (kMER) [247]. In the kMER algorithm, each neuron i, with weight vector w_i is activated by data point V when $|| \ v - w_i \ ||< \sigma_i$, where σ_i is the radius of neuron i. This radius defines a hyperspherical region S_i which intersects at threshold τ_i with the Receptive Field (RF) K, where K is usually defined as a unit-height Gaussian centred at w_i. When a data point falls within S_i, the neuron is activated, the weights are updated according to a neighbourhood function λ, and the threshold is raised, otherwise the threshold is lowered. In this sense, the weights, as in a SOM, are adapted to produce a topology-preserving map. The equiprobabilistic aspect of the mapping is then achieved by updating the radii according to the local input density in such a way that, at convergence, the probability of neuron i being activated is given by $\frac{\rho}{N}$, where ρ is a scale factor. The difference from the SOM algorithm can therefore be summarised as follows: the SOM converges towards a mapping which will minimise the distortion (mean squared error) associated with quantizing the input space into non-overlapping Voronoi spaces, while the kMER

algorithm converges towards a mapping which maximises the entropy associated with quantizing the input space into overlapping receptive fields. The authors in [248] also argue that, since the kMER algorithm will result in a more equitable weight distribution, it will produce a better density estimate.

In the next chapter we will apply the infrastructure and algorithms discussed thus far to an appropriately 'difficult' problem in the domain of neuroscience in order to demonstrate their applicability and functionality.

CHAPTER 4

The Role of Tbx1 in Adult Neurogenesis, Schizophrenia, and a 22q11.2-Associated Mouse Model of Autism Spectrum Disorder

The Tbx1 ChIP data and ultrasonic mouse vocalisations used in this chapter were provided by the lab of Dr. Noboru Hiroi, Albert Einstein College of Medicine. All subsequent ChIP-seq, call sequence, entropy, and classification analyses, as well as all literature and resource mining were carried out by the candidate.

4.1 Introduction

22q11.2 deletion syndrome (22q11DS), including DiGeorge syndrome (DGS) [249], velocardiofacial syndrome (VCFS) [250], and conotruncal anomaly face syndrome (CTAF) [251] (amongst others), is an autosomal dominant disorder with an estimated prevalence of approximately one in 4,000 births [252] making it the most common microdeletion syndrome in humans [253]. Typically associated clinical findings of 22q11DS can be broadly separated into two classes. The first of these relates to structural defects, with the affected structures being derived from the pharyngeal apparatus during early embryo development; outcomes include congenital heart disease, palatal abnormalities, characteristic facial features, hypocalcemia due to hypoparathyroidism, and T-cell immunodeficiency as a result of thymic hypoplasia [254, 255, 256]. The second class of findings are more cognitive in nature and include learning difficulties, developmental delay in both motor skills and language emergence, and associated problems in working memory, visual-spatial processing, and non-verbal communication skills [257, 258, 259]. Deletions in the 22q11.2 region have also been linked to a number of neuro-behavioural and psychiatric illnesses including attention deficit hyperactivity disorder (ADHD), autism spectrum disorder (ASD), anxiety, depression, bipolar disorder, schizophrenia, and schizoaffective disorders [260, 261, 262]. In particular, Bassett et al. [263] showed that microdeletions

in 22q11.2 are associated with a 20- to 30-fold increased risk of schizophrenia, with approximately 25% of cases developing the disease [264]. ASD is also strikingly prevalent – in a study of 98 children with a confirmed 22q11.2 deletion, Fine et al. [265] reported ASD symptoms in over 20% of cases, with ∼14% qualifying for a clinical diagnosis of ASD based on the Autism Diagnostic Interview-Revised (ADI-R), a structured interview for parents of possible ASD cases carried out by a specifically trained psychologist. A similar study conducted by the authors in [266] with 60 patients between the ages of 9 and 18 years demonstrated a 50% rate of ASD, with 27 cases being diagnosed as pervasive development disorder-not otherwise specified (PDD-NOS), one of the autism spectrum disorders. In total, more than two-thirds of the patients in the study were classified as having one or more psychiatric disorders according to DSM-IV criteria.

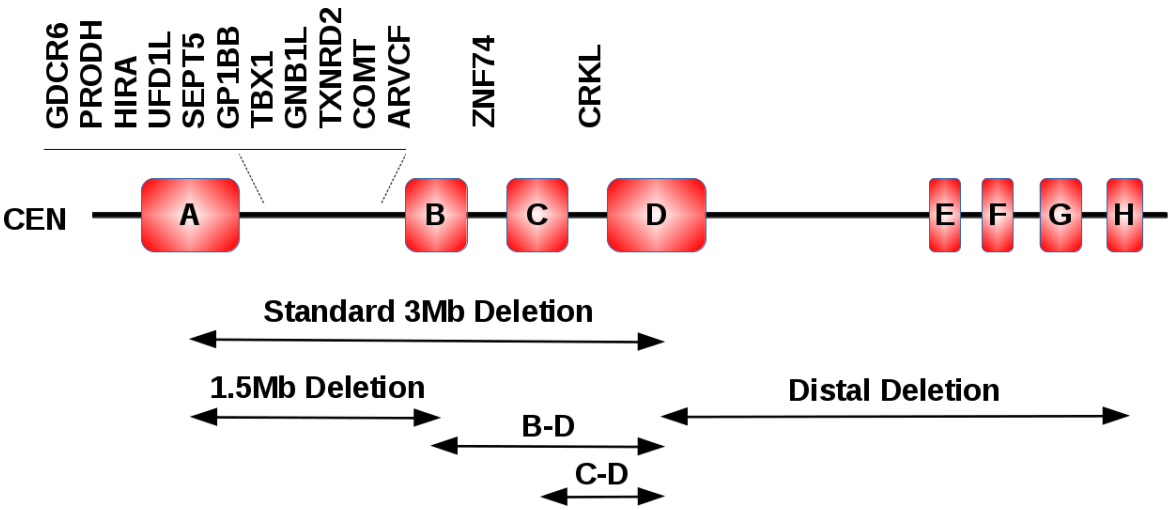

Figure 4.1: 22q11.2 Deletion Syndrome Several microdeletions are associated with 22q11DS, the most common of which is a 3 Mb typically deleted region (TRD) present in ∼85% of cases. The second most common deleteion, ocurring in ∼8% of cases, is a 1.5 Mb region nested within the TDR. Other less common nested deletions involve regions B-D and C-D. More rarely, distal deletions can also occur. (CEN, centromere).

As indicated by its name, 22q11.2 deletion syndrome is caused by microdeletions in the q11.2 region of chromosome 22; the most common of these (present in ∼85% of cases) is a 3 Mb deletion known as the typically deleted region (TDR) which includes 30-45 genes [250], some of which are shown in Figure 4.1. In 90-95% of cases, these microdeletions occur *de novo* and are thought to be caused by non-allelic homologous recombination (NAHR) mediated by the presence of low-copy repeats (LCRs), or segmental duplications, flanking the TDR [253]. Early work in mouse models to determine which of the genes in the TDR might be responsible for 22q11DS phenotypes focused on targeted deletions of a region on mouse chromosome 16 which was shown to be orthologous to

human 22q11.2 [267, 268]. It was demonstrated that deletion of this region on one copy of mouse chromosome 16 produced defects similar to those seen in 22q11DS and that the normal phenotype could be rescued in mice that carried a corresponding duplication on the other chromosome designed to restore the normal gene dosage [269]. Successive experiments with a partial overlap of this deletion showed normal cardiac development [270], narrowing the number of implicated genes, and eventually, individual gene knockout experiments by several independent groups [271, 272] verified that T-box 1 (Tbx1) haploinsufficiency was primarily responsible for the cardiac, conotruncal and parathyroid defects, and moreover that Tbx1-null mice encompass almost all of the common 22q11DS features [273]. It was later discovered that some of the variability present in the range of defects and penetrance associated with Tbx1 haploinsufficiency could be explained by the presence of modifiers of Tbx1 function such as Crkl (v-crk avian sarcoma virus CT10 oncogene homolog-like) also being found in the TDR. Guris et al. [274], for example, showed that heterozygosity in both Tbx1 and Crkl resulted in more severe aortic arch, thymic, and parathyroid defects.

TBX1 is a transcription factor from the evolutionarily conserved T-box family, characterised based on their homology to the initially described Brachyury (T) protein [275]. T-box proteins play an important role during embryogenesis with each member of the family showing highly specific spatiotemporal patterns of expression [276]. TBX1 is expressed in the pharyngeal apparatus [273, 277] where it interacts with members of the fibroblast growth factor (FGF) signalling pathway [278], specifically FGF8 and FGF10. It has been demonstrated that TBX1, through regulation of the gastrulation brain homeobox 2 (GBX2) transcription factor [279], controls the migration of cardiac neural crest cell (cNCCs), a subgroup of multipotent neural crest cells which migrate from the developing neural tube into the pharyngeal arches during pharyngeal arch artery (PAA) development. These cNCCs form the smooth muscle wall of the PAAs and are also involved in the formation of the connective tissue of the thyroid, parathyroid, and thymus glands [280, 281]. TBX1 has also been linked to other key developmental signalling pathways including retinoic acid (RA) signalling [282] (which controls HOX gene expression [283] and is essential for vertebrate organogenesis [284, 285]), and sonic hedgehog (SHH) signalling [286] (also expressed in pharyngeal arches and has been shown to regulate TBX1 via enhancer binding of intermediary FOX transcription factors [287]).

Aside from the obvious structural defects related to Tbx1 haploinsufficiency, Paylor et al. [288] have demonstrated that Tbx1 mutant mice show impaired prepulse inhibition (PPI), a behavioural abnormality associated with several psychiatric disorders including schizophrenia. They also identify a TBX1 frameshift mutation in a family without the common 22q11.2 microdeletions but who nevertheless present with many of the 22q11DS features, including one member with Asperger syndrome. Furthermore, our collaborators in the Hiroi lab (Departments of Psychiatry and Behavioral Sciences, Neuroscience, and Genetics, Albert Einstein College of Medicine) have recently demonstrated that two month old congenic Tbx1 heterozygous mice display schizophrenia and ASD-related behavioural phenotypes including the characteristic ASD features of impaired social interaction and communication [289]. This study provides evidence that frequency and range of ultrasonic vocalisation in mouse pups emitted as a result of maternal separation showed distinct differences between wildtype (WT) and heterozygous (HT) mice, with HT mice demonstrating a more limited usage of the range

of call types studied. The study also indicated that Tbx1, while expressed throughout the brain, was enriched in postnatally generated cells. Postnatal (or adult) neurogenesis primarily occurs in both the subventricular zone (SVZ) which lines the lateral ventricles (LVs) and in the subgranular zone (SGZ) of the dentate gyrus (DG) in the hippocampus (Figure 4.2). In rodents, neuroblasts derived from self-renewing, multipotent, neural stem cell (NSCs) in the SVZ migrate to the olfactory bulb (OB) via the rostral migratory stream (RMS) (Figure 4.3) where they differentiate to form inhibitory interneurons essential for the proper integration of new neurons [290].

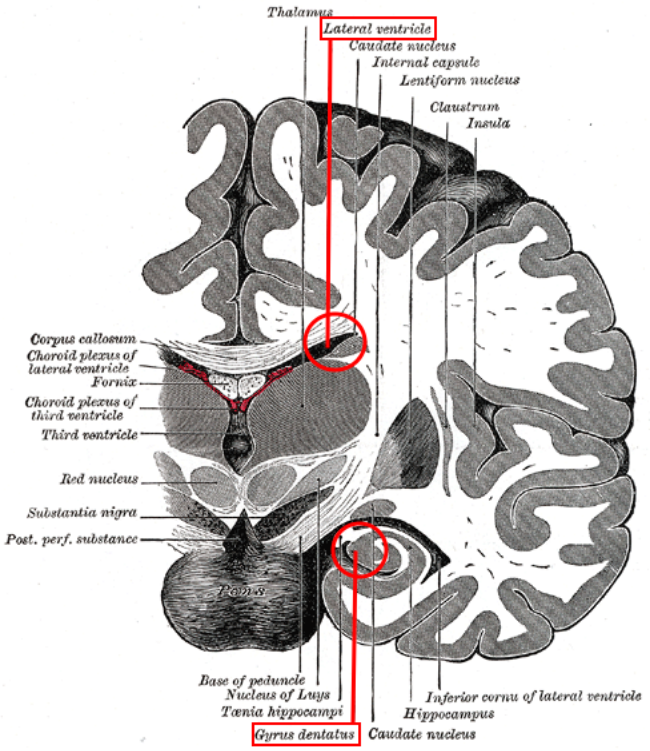

Figure 4.2: Areas of Adult Neurogenesis The two areas of adult neurogenesis in the human brain are shown. The subventricular zone lines the lateral ventricles while the subgranular zone lies in the dentate gyrus of the hippocampus. (Source: Wikimedia Commons, public domain.)

The existence of a human RMS has been proposed [292] but remains controversial [293] despite the fact that migration of neuroblasts from the SVZ to the OB has been observed in human infants, pointing to an important role in synaptic development [290]. It has also been shown that following injuries such as stroke, SVZ-derived neuroblasts can migrate to areas of injury and neurodegeneration [294]. In contrast, NSCs and neural progenitor cells (NPCs) in the dentate gyrus produce neuroblasts which develop into excitatory mature granule cells which are then locally integrated into the granule cell layer and have been linked to processes involved in learning and episodic memory [295, 296].

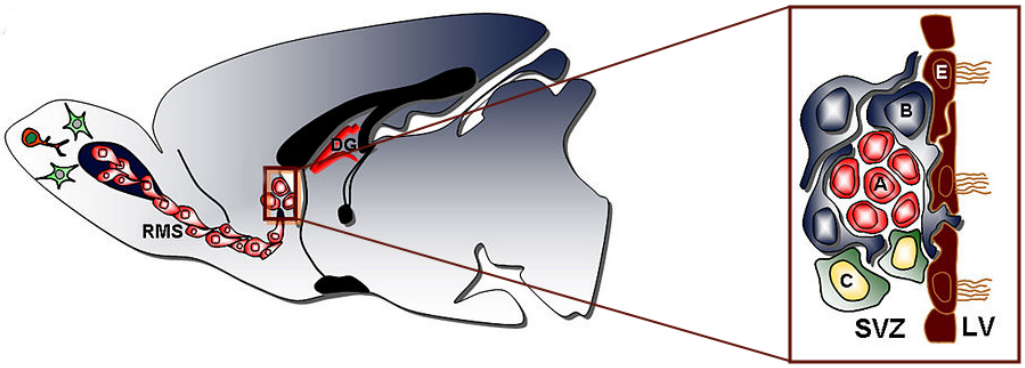

Figure 4.3: Rostral Migratory Stream Main panel shows neuroblasts derived from the rodent SVZ migrating along the rostral migratory stream (RMS) to the olfactory bulb. Insert depicts detailed structure of the rodent subventricular zone (SVZ) including: neural stem cells (A), astrocytes (B), transit amplifying cells (C, more differentiated than neural stem cells), and ependymal cells (E), which line the ventricles and produce cerebrospinal fluid (CSF). DG: dentate gyrus; LV: lateral ventricle. (Source: Wikimedia Commons. Adapted from [291] (C) 2008 Arias-Carrin and licenced under CC-BY-2.0)

These observations by the Hiroi lab regarding Tbx1 expression in postnatally proliferating cells and the resulting schizophrenia and ASD-like behavioural phenotypes observed at two months led to the development of the current study, which comprises two distinct yet complementary parts as will be described in detail the next two sections. Briefly, the first part of this study aims to investigate the genome-wide binding of Tbx1 in postnatal neural progenitor cells in order to identify potential regulatory targets and to examine the role that those targets might play in ASD, schizophrenia, and adult neurogenesis. The second component of the study aims to further investigate differences in social communication in Tbx1 knockdown mice (which exhibit the ASD-like phenotype) and their wildtype counterparts. These two complementary studies therefore aim to link alterations in gene regulation in the developing brain to phenotypic and behavioural differences, providing initial insights into potentially responsible genetic mechanisms.

4.2 Genome-Wide Binding of Tbx1 in Postnatal Neural Progenitor Cells

In order to further examine the role of Tbx1 in adult neurogenesis and to explore any links with schizophrenia and ASD, the first part of this study involved the genome-wide identification of potential Tbx1 targets in murine postnatal neural progenitor cells. Postnatal day-zero (P0) cells derived from the dentate gyrus were chosen for several reasons. Firstly, the hippocampal region has been shown to be involved in both social interaction [297] and PPI [298]. Secondly, day-zero cells were chosen based on the fact that, 1) expression data from [289] indicates that Tbx1 expression was

relatively low at embryonic timepoints compared to the period after birth and that Tbx1 was enriched in postnatally proliferating cells, and 2) evidence presented by [299] indicates that granule cells born during P0-2 number significantly higher in the adult dentate gyrus than cells born at later timepoints. As it is estimated that it takes approximately eight weeks for NPCs to reach maturity in the dentate gyrus [300], we can expect any effects of their altered proliferation at P0 to become apparent at around two months of age, which is when distinct behavioural phenotypes were observed by Hiramoto et al. [289].

4.2.1 Primary Analysis using WASP

Sequencing and Peak Annotation

Neural progenitor cell cultures derived from the hippocampal dentate gyrus of postnatal day-zero (P0) C57BL/6J pups were first treated with formaldehyde to cross-link DNA and proteins. Samples were then sonicated to shear chromatin into fragments with an average size of 200-500 bp. Five percent of the pre-immunoprecipitated solution was saved as input DNA, while an antibody specific to Tbx1 (Abcam) was added to the remainder of the sample. The chromatin-antibody complexes were precipitated using Invitrogen Dynabeads protein G and the concentration and quality of released and purified DNA was determined using an Agilent 2100 Bioanalyzer. Enrichment was assessed by amplifying input and ChIP DNA with primers specific to known positive and negative control loci. Following QC, Illumina adapters were ligated to both input and IP samples and the resulting libraries were sequenced using the GAIIx system. Data generated by the sequencer was automatically processed by the WASP ChIP-seq analysis pipeline which includes read quality analysis, alignment to the appropriate reference genome using Bowtie, and peak-calling using MACS with an m-fold parameter of 10-30 and a p-value cut-off of 1e-05.

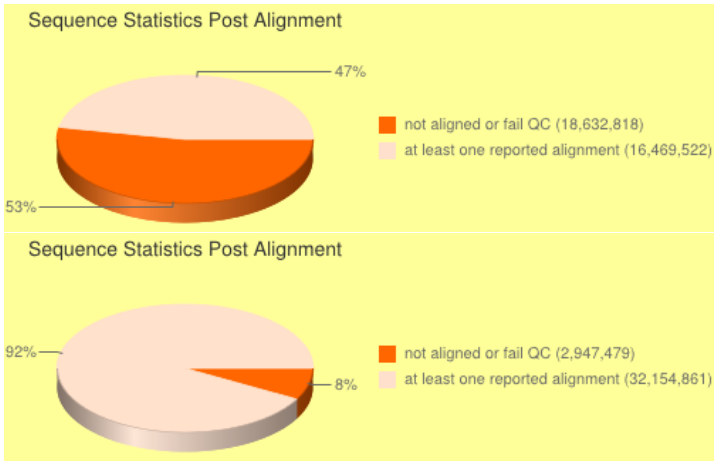

Figure 4.4: Read Trimming Initial alignment (upper panel) resulted in less than half of the generated reads mapping to the reference genome. Following read trimming, this number increased to over 90% (lower panel).

Upon initial examination of the MACS ouput for this study we noticed a relatively small number of called peaks (325). Looking at the quality plots for the run, we identified this issue as stemming from the fact that less than half of the generated reads (∼47%) were actually aligning to the genome (Figure 4.4). Further examination of WASP output for assessing read quality and composition showed an unusual pattern of k-mer enrichment in the end of the reads possibly indicating a problem with library construction. We therefore used the Sickle[1] tool to trim the reads and proceeded to re-run the ChIP-seq pipeline on these shorter reads. This resulted in both a large increase in the percentage of aligned reads (∼92%) and in a subsequent increase in the number of MACS-called peaks (2,374).

In order to assign annotations to the identified ChIP peaks, the GREAT tool [301] was used. Specifically, a basal regulatory domain of 5kb upstream of TSS and 1kb downstream of TSS was defined for each anotated gene in the mm9 genome. This regulatory domain is extended in both directions to provide a distal regulatory domain, with the extension proceeding only as far as the nearest neighbouring gene's basal domain, or alternatively, to a predefined maximum of 100kb. Peaks falling within a gene's proximal or distal regulatory domain were assigned to that gene, with the expectation being that they may help to regulate expression of the gene through either promoter or enhancer binding. Of the 2,374 peaks, 1,587 did not map to the regulatory domain of any gene, 524 mapped to the regulatory domain of one gene, while 263 mapped to the regulatory domain of two distinct genes (Figure 4.5). The 1,587 peaks not mapping to any gene may still in fact bind to enhancer elements, but this is difficult to assess given that the analysis does not account for extremely distal elements being brought into contact with more proximal regulatory regions through chromatin looping. Figure 4.5 also shows the distance histogram for the 1,050 peak-gene associations – 53 of these peaks map within 5kb of their associated gene's transcription start site, 540 map within 50kb, and 457 map within 100kb. In total, accounting for the fact that multiple peaks may map to the same gene, 407 unique genes are identified as potential Tbx1 targets.

Autism and Schizophrenia Related Genes

In order to ascertain if any of our potential Tbx1 targets are ASD related, we used two publicly available resources – AutismKB [302] and SFARI[2]. SFARI is a licensed copy of AutDB [303], a curated, web-based, resource which includes information on human genes, animal models, protein interactions, and CNVs, while AutismKB is an evidence-based repository for autism related genes and CNVs, with data curated from over 616 published studies, including GWAS, expression profiling, low throughput genetic association, and CNV experiments. Each gene in the AutismKB repository is assigned a score based on the level of evidence for its association with autism from across the mined publications, and in total, 99 syndromic genes, 3,050 non-syndromic genes, and more than 4,500 CNVs are included. A high-confidence 'core' subset of 171 autism-candidate genes is also defined based on combined evidence. Similarly, we use two popular schizophrenia resources to define any

[1] http://github.com/najoshi/sickle/
[2] http://sfari.org/

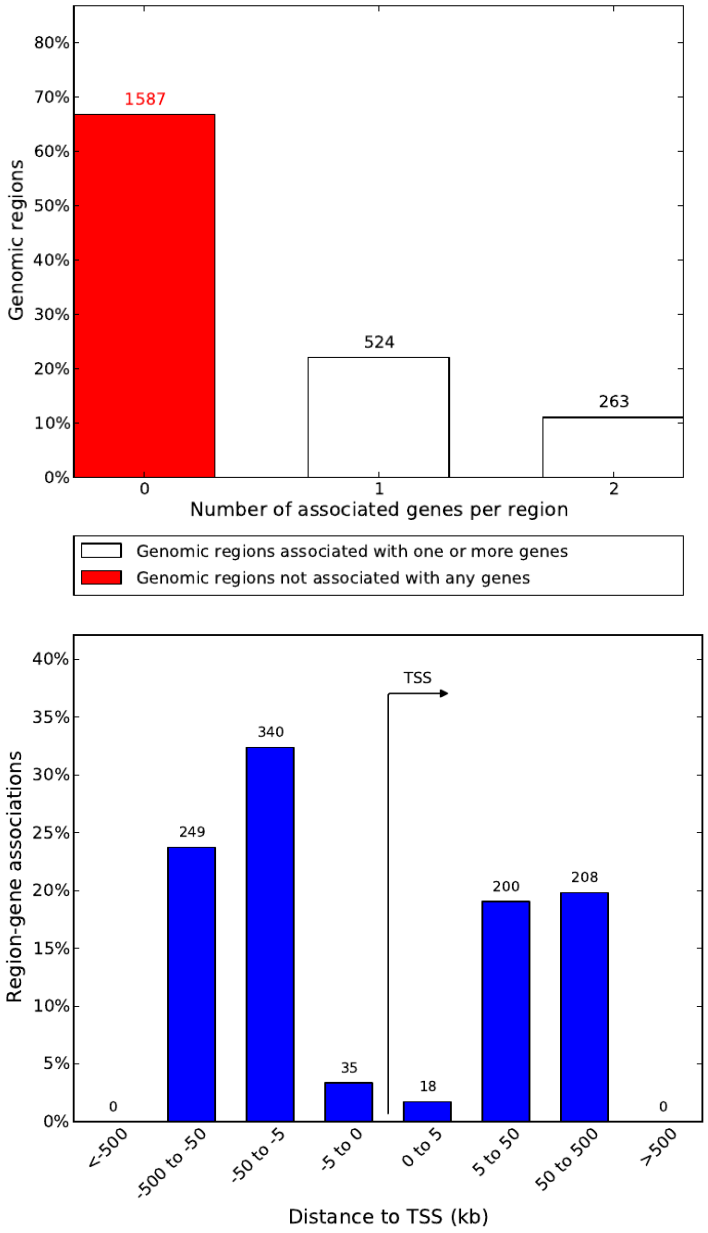

Figure 4.5: Peak Annotation Upper panel shows the number of ChIP-seq peaks mapping to zero, one, or two gene annotations using the GREAT tool. The lower panel shows the distribution of distances to TSS for each peak-annotation association.

overlap of our peak list with known or predicted schizophrenia related genes. Szgene [304] contains 1,008 candidate schizophrenia genes derived from over 1700 genetic association studies, while the Schizophrenia Gene Resource (SZGR) [305] provides a list of 75 prioritised genes derived using a combined odds ratio (COR) approach [306] on 500 genes mined from over 2,000 association studies.

Our cohort of 407 Tbx1-bound genes contains a total of 51 genes from AutismKB. Five of these genes are syndromic, nine are from the high-confidence subset, and six have a Tbx1 peak within 5kb of their TSS (Table 4.1). Of the nine high-confidence hits, seven (AUTS2, DISC1, EN2, FBXO33, PTEN, SCN1A, and TBR1) are also matched when comparing our list of peaks to the SFARI database. Three further genes (DYRK1A, IL1R2, and SLC38A10) are matched to SFARI candidates with two of these (DYRK1A and IL1R2) being found in the wider list of 51 hits at AutismKB. Of the seven genes matched by both AutismKB and SFARI, two (DISC1 and PTEN) appear in SZGene (out of a total of 17 matches), but while DISC1 also appears in SZGR, PTEN does not. DISC1 is also listed (along with TBR1) at MANGO, the mammalian adult neurogenesis gene gntology [307]. This ontology lists genes which are important for different stages of hippocampal neurogenesis, from proliferation, to differentiation, migration, neuritogenesis, and survival. Below, we provide details on these seven high-confidence hits and further examine their respective roles in autism, schizophrenia and neurogenesis.

AUTS2 Autism suceptibility candidate 2 (reviewed in [308]) was first linked to autism in [309] when the authors, studying a pair of monozygotic twins with ASD, identified it as a balanced t(7;20)(q11.2; p11.2) translocation breakpoint. It has since been linked to schizoaffective [310] and bipolar disorders [311], delayed language development [312], dyslexia [313], ADHD [314], and epilepsy[315]. In terms of neurogenesis, its expression has been shown in multiple regions of the brain, including in the dentate gyrus of human fetal brains at 23 weeks [316]. In [317], the authors more fully explored AUTS2 expression in the developing mouse brain from E11 to P21. They showed that it co-localises early in the process with TBR1, which regulates its expression, that it is expressed at different levels and in different regions throughout the course of brain development, and that postnatally (P21), it was found to be expressed in the hippocampus throughout the subgranular zone and granule cell layer.

DISC1 Disrupted in schizophrenia 1 was first identified in 1990 in a Scottish family affected with a range of psychiatric illnesses including schizophrenia, bipolar disorder, and major depression [318]. It is part of a balanced t(1;11)(q42.1;q14.3) translocation which also disrupts the DISC2 gene, an antisense noncoding RNA, which may play a role in the regulation of DISC1 [319]. It is a key regulator of both embryonic and adult neurogenesis [320] playing a role in cell proliferation, differentiation and migration [321]. In [322], the authors show that it is both highly expressed in the embryonic subventricular zone and that its knockdown results in decreased proliferation of adult neural progenitor cells in the dentate gyrus. They identify DISC1's interaction with the WNT pathway as a mechanism for this effect, demonstrating that its knockdown results in decreased levels of β-catenin and disrupts the ability of the int/Wingless family member WNT3A to stimulate progenitor cell proliferation. DISC1 has been shown to control the tempo of neuronal integration in the adult hippocampus through regulation of the AKT-mTOR pathway [321] with downregulation resulting in an acceleration in integration leading to aberrant morphology and mispositioning of newly generated granule cells in the dentate gyrus [323]. In addition to its role in schizophrenia and

associated affective disorders, DISC1 has also recently been linked to both autism [324, 325] and Asperger syndrome [325].

EN2 Engrailed 2 is a homeobox transcription factor that is important in multiple aspects of brain development, including the regulation of midbrain and hindbrain development [326] and pattering of the cerebellum [327]. Postnatally, EN2 is primarily expressed in mature granule cells in the cerebellum [328], but has also recently been shown to be present at lower levels in the hippocampus and cerebral cortex [329]. EN2 knockout mice display a reduction in Purkinje cells (which has also been observed in schizophrenic and bipolar individuals [330]) and changes in social and motor behaviour similar to ASD [331]. The authors in [332] show a reduction in hippocampal weight in P21 EN2 null mice, as well as an increased turnover in cells of both the dentate gyrus and subventricular zone. In humans, two intronic SNPs in EN2 (rs1861972 and rs1861973) have been shown to be associated with ASD [333] with the AC haplotype overrepresented in affected individuals. Other studies have confirmed this association [334, 335] and also linked SNPs in EN2 with young-onset Parkinson's disease [336].

FBXO33 F-Box protein 33, like other F-box proteins is an adaptor protein which can form part of the Skp-Cullin-F-box (SCF) complex, an E3 ubiquitin ligase. The F-box component of this complex is responsible for targetting specific protein substrates for ubiquitination and eventual degradation by the proteasome [337]. A recent publication by Glessner et al. [338] has demonstrated an association between CNVs involving genes playing a role in the ubiquitin pathway and ASD. One of the novel CNVs presented in this study includes another F-box protein, FBXO40. FBXO33 itself also lies in a region (14q21.1) which has been identified in multiple studies as a risk locus for ASD [339, 340] and developmental delays [341]. Along with NXF, FBXO33 was shown in mouse studies to be upregulated in the hippocampus one hour after pharmacologically induced seizure [342].

PTEN Phosphatase and tensin homolog is a protein and lipid phosphatase that functions as a tumor supressor through its regulation of the AKT-mTOR (mammalian target of rapamycin) pathway via inhibition of phosphoinositide 3-kinase (PI3K) [343]. It has been shown to be mutated in a large number of cancers including glioblastoma, prostate, and breast cancer [344, 345, 346] and is also responsible for several multiple hamartoma syndromes including Cowden's disease [347] and Bannayan-Riley-Ruvalcaba syndrome [348]. PTEN mutations have also been linked to autism, developmental delays, and macrocephaly [349, 350, 351]. Kwon et al. [352] have demonstrated that deletion of PTEN in differentiated neuronal populations of the cerebral cortex and hippocampus results in abnormal dendritic and axonal growth and an ASD-like social and behavioural phenotype. PTEN has been shown to play a role in the regulation of both embryonic [353] as well as adult [354] neural stem cells, and this regulation is evident in both the SVZ and the hippocampus. The authors in [355], for example, demonstrate that PTEN deletion in postnatal hippocampal NSCs results in increased proliferation and differentiation leading to macrocephaly with an enlarged dentate gyrus, early depletion of the NSC pool, and impairment in social interaction.

SCN1A Sodium channel, voltage-gated, type I (alpha subunit) is another of the syndromic genes on our high-confidence list. Mutations in this gene are associated with a host of seizure related disorders (a review of some 60 different frameshift, missense and nonsense mutations can be found in [356]), ranging from the milder febrile seizures (FS), to the more severe intractable childhood epilepsy with generalised tonic-clonic seizures (ICE-GTC) [357] and Dravet Syndrome [358] (also known as severe myoclonic epilepsy in infancy SMEI), or polymorphic myoclonic epilepsy in infancy (PMEI)). These seizures usually begin within the first year, are characterised as unusually severe, and in the more serious cases of ICE-GTC and Dravet syndrome, can cause mild to severe cognitive impairments, developmental delay, behavioural disturbances, and psychomotor dysfunction [359, 360]. Mutations in SCN1A and its family member SCN2A have also recently been linked to ASD [361, 362].

TBR1 T-box brain 1, a homolog of Brachyury [363], is a brain-expressed T-box protein which plays a role in neuronal differentiation [364] and aids in neuronal migration and axon guidance through regulation of Reln [365], a gene shown to be expressed at decreased levels in individuals with ASD [366]. It has also been shown to regulate another key candidate gene from our list, AUTS2, during mouse brain development [317]. A recent study by Roybon et al. has shown that, aside from its role in early brain development, TBR1 (along with its family member TBR2) is expressed in a population of NSCs and progenitor cells found in the SVZ-RMS axis, indicating a role for these transcription factors in adult neurogenesis in the olfactory bulb [367].

Based on Table 4.1, DISC1 and PTEN were selected as initial validation targets as they show the largest overlap in terms of matches at ASD, schizophrenia, and neurogenesis resources. Our collaborators in the Hiroi lab have started this validation process and have so far shown using a luciferase reporter assay coupled with a Tbx1-specific siRNA and qRT-PCR, that Tbx1 does in fact bind to the upstream promoter region of Pten and drives its expression. At the time of writing, validation of Disc1 as well as further high-priority targets is currently under way.

While we have thus far focused primarily on seven genes which show the highest overlap between matches in both autism databases, there are many other genes found in our complete list of Tbx1 peak associations that are not part of this high-confidence subset which nevertheless show strong associations with both psychiatric illnesses, and more generally with the process of neurogenesis. Two other genes, for example, listed at MANGO also appear on our peak list – secreted phosphoprotein 1 (SPP1, also matched at SZGene) has been shown to play a role in the migration of neuroblasts in response to cerebral ischemia [368], while GRIA1, a glutamate receptor, (also listed at SZGene and AutismKB) has been linked to a number of diseases from schizophrenia [369], to Alzheimers disease [370], bipolar disorder [371], and epilepsy [372]. Similarly, DDX26, a DEAD box protein has previously been linked to both autism and schizophrenia [373], with the authors in [374] using it as one of 14 predictive genes in an SVM-based classifier to segregate subtypes of severely language-impaired ASD patients from corresponding controls with sensitivity and specificity levels both greater than 90%.

Gene Symbol	Gene Name	Cytoband	OMIM ID	HC	PP	SF	SG	SZ
ACACB	acetyl-CoA carboxylase beta	12q24.11						
ACP6	acid phosphatase 6, lysophosphatidic	1q21						
AMY2B	amylase, alpha 2B	1p21						
APAF1	apoptotic peptidase activating factor 1	12q23						
ARNT	aryl hydrocarbon receptor nuclear translocator	1q21						
ARRDC4	arrestin domain containing 4	15q26.3						
AUTS2	autism susceptibility candidate 2	7q11.22		✓		✓		
BRD3	bromodomain containing 3	9q34						
COL4A1	collagen, type IV, alpha 1	13q34						
CR2	complement component (3d/Epstein Barr virus) receptor 2	1q32						
CSDA	cold shock domain protein A	12p13.1						
CYR61	cysteine-rich, angiogenic inducer, 61	1p22.3						
DCX	doublecortin	Xq22.3-q23	300067	✓				
DISC1	disrupted in schizophrenia 1	1q42.1		✓	✓	✓	✓	✓
DYRK1A	dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 1A	21q22.13				✓		
EN2	engrailed homeobox 2	7q36		✓		✓		
EPC1	enhancer of polycomb homolog 1	10p11						
FAR2	fatty acyl CoA reductase 2	12p11.22						
FBXO33	F-box protein 33	14q21.1		✓		✓		
FGFR2	fibroblast growth factor receptor 2	10q26	101200				✓	
GAD2	glutamate decarboxylase 2	10p11.23					✓	
GC	group-specific component	4q12-q13					✓	
GRIA1	glutamate receptor, ionotropic, AMPA 1	5q33—5q31.1					✓	
IL1R2	interleukin 1 receptor, type II	2q12				✓		
ITGA11	integrin, alpha 11	15q23						
KLHL8	kelch-like 8	4q22.1						
MUM1L1	melanoma associated antigen (mutated) 1-like 1	Xq22.3						
MYO5B	myosin VB	18q21						
MYO7A	myosin VIIA	11q13.5						
NCALD	neurocalcin delta	8q22.2						
NGFRAP1	nerve growth factor receptor associated protein 1	Xq22.2						
NPFFR2	neuropeptide FF receptor 2	4q21						
PDE7B	phosphodiesterase 7B	6q23-q24					✓	
PDIA6	protein disulfide isomerase family A, member 6	2p25.1						

continued

Gene Symbol	Gene Name	Cytoband	OMIM ID	HC	PP	SF	SG	SZ
PELI3	pellino homolog 3	11q13.2			✓			
PTEN	phosphatase and tensin homolog	10q23.3	601728	✓	✓	✓	✓	
RASIP1	Ras interacting protein 1	19q13.33			✓			
RNF135	ring finger protein 135	17q11.2	611358	✓	✓			
RORB	RAR-related orphan receptor B	9q22					✓	
RPS29	ribosomal protein S29	14q			✓			
SCN1A	sodium channel, voltage-gated, type I	2q24.3	607208	✓		✓		
SLC20A2	solute carrier family 20	8p12-p11						
SP7	Sp7 transcription factor	12q13.13						
SULF1	sulfatase 1	8q13.1						
TBC1D23	TBC1 domain family, member 23	3q12.2						
TBR1	T-box, brain, 1	2q24		✓		✓		
TBX21	T-box 21	17q21.32						
TGM3	transglutaminase 3	20q11.2						
TNFRSF8	tumor necrosis factor receptor superfamily, member 8	1p36						
TRO	trophinin	Xp11.22-p11.21						
WHAMM	WAS protein homolog	15q25.2						

Table 4.1: AutismKB Genes Genes listed at AutismKB which contain a Tbx1 binding peak. Gene symbols, names and genomic locations are shown, as are associated syndromes (OMIM IDs), membership of the high-confidence 'core' subset at AutismKB (HC), proximal peak (PP – defined as within 5Kb of TSS), and overlap with SFARI (SF), SZGene (SG), and SZGR (SZ) resources.

Autism and Schizophrenia Related CNVs

We were also interested in examining if any of the other 407 potentially Tbx1-bound genes from our peak list which did not appear in the gene-centric resources mentioned in the previous section might lie in genomic regions with known structural variants associated with either ASD or schizophrenia – any genes identified as such may represent novel candidates worth confirmatory study. We therefore extracted all known genes lying in regions defined by the more than 4,500 CNVs in AutismKB, as well as those found in 27 schizophrenia related CNVs from a recent publication by Rees et al. [375] which included a review of 13 previously defined loci as well as providing 12 novel ones. Of the 407 genes identified, 128 had genomic coordinates which placed them in AutismKB listed CNVs. Of these, 104 (Table 4.3) were not listed as part of the cohort of 51 AutismKB genes shown in Table 4.1, indicating that this may indeed be a useful approach to identifying currently unverified candidate genes. Only three genes from our peak list (ACP6, GJA5, BDH1) were identified in the 27 schizophrenia related CNVs (Table 4.2). All three are also present on the Autism CNV list, again highlighting the genetic links between these two disorders.

Gene Symbol	Gene Name	Cytoband
ACP6	acid phosphatase 6	1q21.2
GJA5	gap junction membrane channel protein alpha 5	1q21.2
BDH1	3-hydroxybutyrate dehydrogenase	3q29

Table 4.2: Schizophrenia CNV Genes These three potentially Tbx1-bound genes are found at genomic loci listed by Rees et al. [375] as being associated with schizophrenia. ACP6 is also listed as part of the cohort of 51 genes found in AutismKB (Table 4.1), while all three also appear on the list of ASD-related CNVs in Table 4.2.

Gene Symbol	Cytoband	Gene Symbol	Cytoband	Gene Symbol	Cytoband	Gene Symbol	Cytoband
ACADS	12q24.31	DTWD1	15q21.2	MYL7	7p13	SNX19	11q24.3
ADAM7	8p21.2	DUSP21	Xp11.3	NAP1L2	Xq13.2	SP1	12q13.13
ADAMDEC1	8p21.2	EDF1	9q34.3	NAPEPLD	7q22.1	SPOCK1	5q31.2
ADAMTSL3	15q25.2	ERGIC2	12p11.22	NCR1	19q13.42	SPP1	4q22.1
ADCY8	8q24.22	FNIP2	4q32.1	NEK1	4q33	SPPL3	12q24.31
ADD3	10q25.1	FTSJD2	6p21.2	NUDT11	Xp11.22	SSX9	Xp11.23
ANO3	11p14.2	GCK	7p13	NUDT19	19q13.11	TDRD3	13q21.2
ATF7IP2	16p13.2	GJA5	1q21.2	PCDHB8	5q31.3	TINAG	6p12.1
ATP6V0C	16p13.3	GLB1L2	11q25	PCDHB9	5q31.3	TMED1	19p13.2
ATXN1L	16q22.2	GLB1L3	11q25	PGAM2	7p13	TMEM106B	7p21.3
AURKC	19q13.43	GNAL	18p11.21	PLEKHA5	12p12.3	TMPRSS5	11q23.2
BARX2	11q24.3	HOMER2	15q25.2	PLIN1	15q26.1	TMX3	18q22.1
BBX	3q13.12	HSD17B13	4q22.1	POLM	7p13	TOX3	16q12.1
BDH1	3q29	IL22	12q15	PPID	4q32.1	TPD52	8q21.13
BZW2	7p21.1	KIF7	15q26.1	PTCD3	2p11.2	TRAF2	9q34.3
CACNB4	2q23.3	LCE6A	1q21.3	PTPRQ	12q21.31	TRAM1L1	4q26
CCBL1	9q34.11	LRRC23	12p13.31	RAMP3	7p13	TREM1	6p21.1
CCNH	5q14.3	LRRC8A	9q34.11	RAP2A	13q32.1	TSN	2q14.3
CD36	7q21.11	LY75	2q24.2	RDH16	12q13.3	TYK2	19p13.2
CDC37	19p13.2	MAGED2	Xp11.21	RGS9BP	19q13.11	UBE2E3	2q31.3
CHST7	Xp11.23	MALT1	18q21.32	RPAP3	12q13.11	UBE2F	2q37.3
CHSY3	5q23.3	MAML2	11q21	RPL39L	3q27.3	UNC93A	6q27
CORO2B	15q23	MEPE	4q22.1	SETDB1	1q21.3	VPS41	7p14.1
DDX26B	Xq26.3	MPPED2	11p14.1	SLC38A1	12q13.11	WDR5	9q34.2
DNM2	19p13.2	MRPS10	6p21.1	SMCP	1q21.3	YES1	18p11.32
DPP9	19p13.3	MRPS31	13q14.11	SMTN	22q12.2	ZSWIM6	5q12.1

Table 4.3: AutismKB CNV Genes Genes which are located in regions containing known CNVs listed as ASD-associated at AutismKB. Of the 128 genes matched in total, these 104 do not appear in Table 4.1, indicating that they may be as yet be unconfirmed ASD candidate genes.

Pathway Analysis

Diseases and Disorders	p-value	Molecules
Cardiovascular Disease	7.27E-05 – 1.47E-02	16
Immunological Disease	7.27E-05 – 1.47E-02	18
Inflammatory Disease	7.27E-05 – 1.47E-02	18
Cancer	2.16E-04 – 1.79E-02	20
Organismal Injury and Abnormalities	2.16E-04 – 1.75E-02	29

Physiological System Development and Function	p-value	Molecules
Hematological System Development and Function	6.10E-06 – 1.95E-02	42
Nervous System Development and Function	2.16E-04 – 1.97E-02	21
Tissue Morphology	2.16E-04 – 2.00E-02	18
Connective Tissue Development and Function	4.52E-04 – 2.00E-02	17
Skeletal and Muscular System Development and Function	4.52E-04 – 1.69E-02	16

Table 4.4: Top Biological Functions An IPA analysis of the 407 Tbx1-bound genes indicates statistically significant enrichment in expected pathways and diseases including hematological, cardiovascular, and nervous system development.

As well as examining genes on an individual basis, we also performed a pathway analysis of the 407 Tbx1 target genes using the Ingenuity Pathway Analysis[3] (IPA) tool to ascertain if there was any statistically significant enrichment in known biological process or processes. This analysis reveals a wide range of molecular activity with pathway and functions tagged as significant showing concurrence with what we know about Tbx1 interactions and have thus far discussed in terms of the effects of Tbx1 haploinsufficiency. The top physiological functions, for example, include hematological, connective tissue, nervous, and skeletal and muscular systems development, while the top diseases include cardiovascular and immunological diseases, as well as organismal injury and abnormalities (Table 4.4). Three of the top molecular networks identified were cardiovascular system development and function, developmental disorder, and organ morphology (data not shown).

4.2.2 Secondary Motif Analysis

As stated in the previous chapter, a secondary analysis of sequences under ChIP-seq identified peaks is typically carried out to ascertain whether or not a known motif is present or if indeed a novel motif can be identified. A direct scan of the peaks was first used to search for the consensus T-box half-site sequences GTGXXA (which occurs 1,698 times) and T(G/C)ACAC (which occurs 211 times). These are however relatively short motif sequences and many if not most of these occurrences may be false positives. In order to determine if there was a statistically significant enrichment of this

[3]http://ingenuity.com

motif and to potentially identify any variations thereof, both CudaMEME [376] and ChIPSOM were used to perform a *de novo* motif analysis.

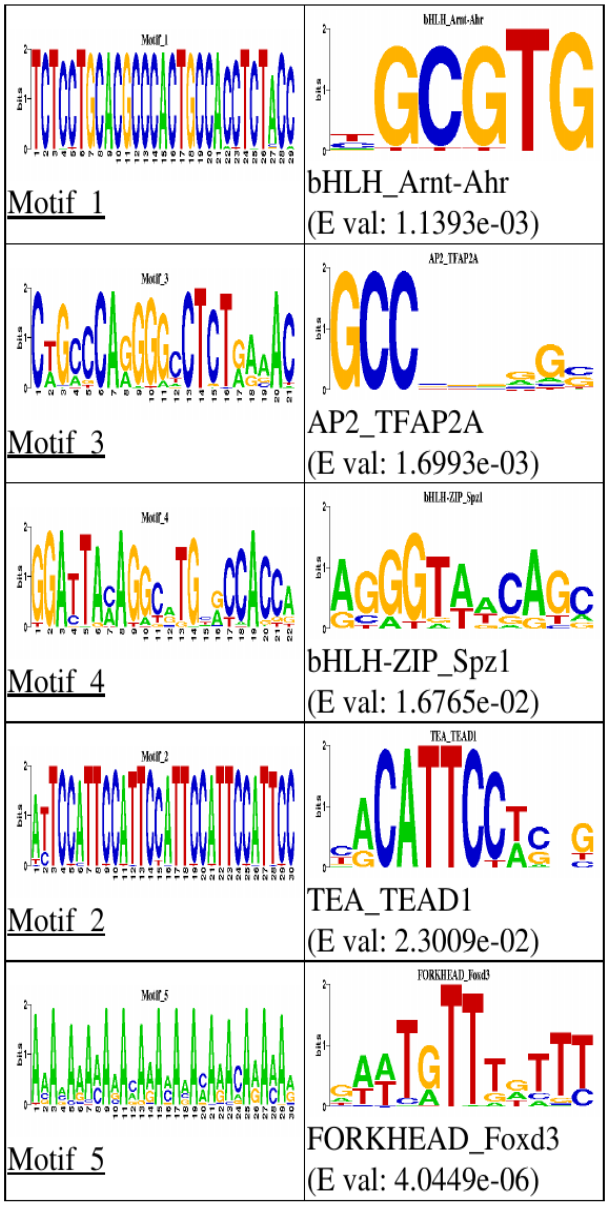

Figure 4.6: CudaMEME Motifs The motifs returned by CudaMEME did not show any matches to T-box motifs in either JASPAR or TRANSFAC. Shown here are the top matches in the JASPAR database, which include bHLH, Forkhead, AP2, and TEA domain motifs.

CudaMEME is a GPU-enabled version of the standard MEME algorithm designed to offer scal-

able performance on large sequence datasets by leveraging the power of general purpose GPU (GP-GPU) cards such as the NVidia Tesla. As discussed in the previous chapter, ChIPSOM is a variation on the original SOMBRERO algorithm designed to allow the processing of larger datasets in a more efficient manner. Both algorithms were set to search for motifs in the 6-30bp range, on both the forward and reverse strands. For CudaMEME, the 'anr' model was used which allows for any number of motif occurrences per peak. ChIPSOM was run with a map size of 25x50 for 100 iterations. The motifs returned were then used as input to the STAMP platform to look for the top five matches to known motif models in both the JASPAR and TRANSFAC motif databases. CudaMEME motifs showed much higher information content than those returned by ChIPSOM but often consisted of simple repeats (Figure 4.6). The top hits did not show any matches to T-box related motifs but included bHLH, Forkhead, AP2, and TEA domain motifs. ChIPSOM returned 41 motif models which passed the statistical threshold for enrichment. Post-ChIPSOM clustering of these motifs using both GMACS and STAMP resulted in 2 main clusters, with one motif outlier and the remaining models clustering together. Of these 41 models, 14 show T-box motif matches as detailed in Table 4.5. An example match is shown for motif 22 (Figure 4.7) which was matched to known T-box models in both the JASPAR and TRANSFAC databases. In total, T-box matched motif instances were found in 44 of the 407 GREAT-annotated genes (10.8%).

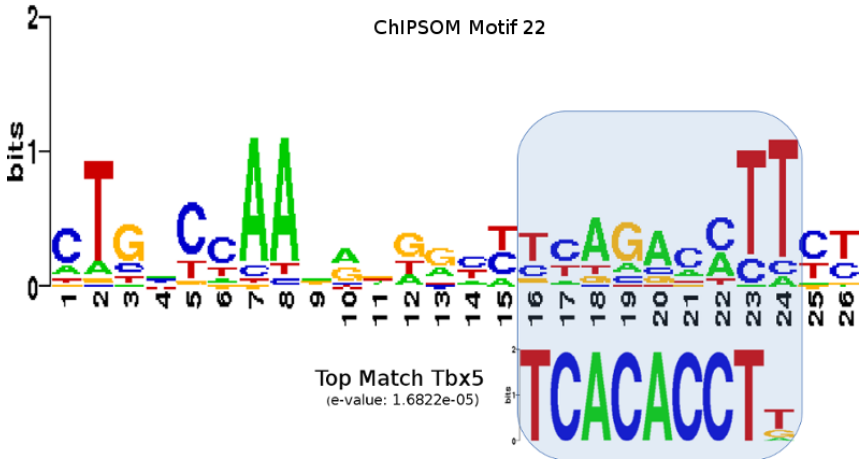

Figure 4.7: ChIPSOM Motif Motif 22 is one of the motif models returned by ChIPSOM which matched known T-box motifs in both the JASPAR and TRANSFAC databases.

Motif	Rank	E-value	DB	MATCH	MATCH CONSENSUS SEQ
22	1	1.68e-05	TF	TBX5	TCACACCTT
27	2	3.71e-04	TF	TBX5	TCACACCTT
34	5	1.64e-03	TF	TBX5	TCACACCTT
33	3	3.05e-04	TF	TBX5	TCACACCTT
38	3	1.25e-04	TF	TBX5	TCACACCTT
13	2	4.52e-04	TF	TBX5	TCACACCTT
35	4	3.45e-04	TF	TBX5	TCACACCTT
39	2	6.98e-04	TF	TBX5	TCACACCTT
22	3	5.08e-03	JASP	T-box-T	TTCACACCTAG
29	2	4.28e-03	JASP	T-box-T	TTCACACCTAG
31	1	1.41e-02	JASP	T-box-T	TTCACACCTAG
25	1	3.45e-03	JASP	T-box-T	TTCACACCTAG
26	3	4.90e-03	JASP	T-box-T	TTCACACCTAG
11	2	3.20e-03	JASP	T-box-T	TTCACACCTAG
9	3	3.17e-03	JASP	T-box-T	TTCACACCTAG

Table 4.5: ChIPSOM T-box Matches Shown in this table are the ChIPSOM motifs which matched known T-box motif models in the JASPAR and TRANSFAC databases. Motif indicates the ChIPSOM motif ID, Rank is the position in the match list (STAMP was set to return the top 5 matches from each of the databases), E-value indicates the likelihood of such a match by chance, DB is the database matched against (TF for TRANSFAC, JASP for JASPAR), the Match column details the motif model matched - all TRANSFAC matches were to the TBX5 model while all JASPAR matches were to the T (Brachyury) motif model.

4.3 Vocalisation in Tbx1 Knockdown Mice

The second part of this study aimed to examine in more detail the differences in social communication in eight different phenotypic groups (MP8WT, MP8HT, FP8WT, FP8HT, MP12WT, MP12HT, FP12WT, FP12HT) incorporating both male (M) and female (F) congenic Tbx1 wild-type (WT) and heterozygous (HT) mice at 8 and 12 days postnatally (P8 and P12 respectively). While the previous study from the Hiroi lab showed in a more limited manner that differences are present in single-call frequencies at P7-8 [289], here, we provide a more rigorous analysis using more sophisticated information theory and machine learning techniques with the goal of identifying both differences in the structure of WT and HT vocalisation patterns, as well as a subset of call sequences which might enable us to readily distinguish WT from HT mice.

4.3.1 Experimental Design

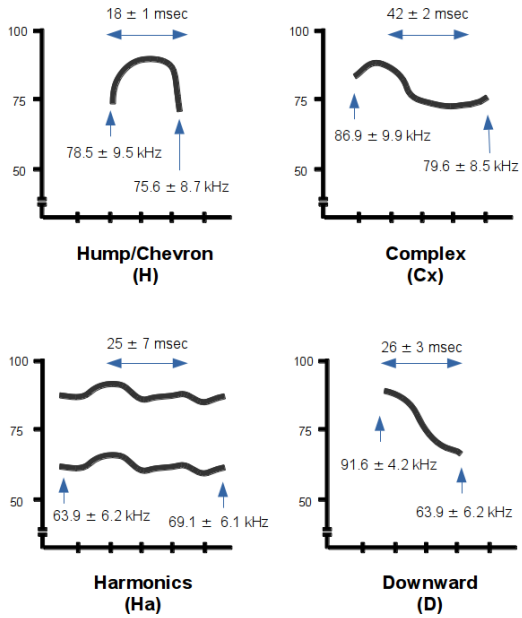

Figure 4.8: Call Spectrograms Example spectrograms are shown for four of the call types used in this study. Spectrograms and values for time and frequency ranges are based on [377]. X-axis represents time in msec, Y-axis is frequency in kHz (data not shown to scale).

Mouse pups were separated from their mothers at either P8 or P12 and placed onto a plastic tray with standard cage bedding. This tray was then placed into a Styrofoam box attached with an Avisoft UltraSoundGate ultrasonic condenser microphone (Avisoft Bioacoustics, Germany) sensitive to frequencies in the 10-200 kHz range. Recordings were made for 5 minutes at a 300kHz sampling rate and spectrograms were automatically produced by the Avisoft-SASLab Pro software using a

fast Fourier transform (FFT). Spectrograms were then interpreted based on 10 distinct call types as outlined in [377] - these call types are: Cx (Complex), Ha (Harmonics), Ts (Two-syllable), U (Upward), D (Downward), H (Hump, also known as Chevron), Sh (Shorts), C (Composite), Fs (Frequency steps), and F (Flat). A cartoon depicting typical examples for some of the spectrograms associated with these call types is shown in Figure 4.8.

These 5 minute call sequences were then analyzed to determine break points between different strings or bursts of vocal emissions. For randomly generated WT and HT sequences, a curve of expected interval frequency was plotted - this curve follows a binomial distribution based on the fact that we are observing a fixed number of calls within a finite timeframe. The observed distributions of intervals from the original WT and HT sequences were overlayed on this plot (Figure 4.9) and the point at which these distributions intersected was taken as the interval length at which to 'cut' the 5 minute sequence. Any calls within the sequence separated by more than this length of interval were considered to represent the ends of two independent call strings.

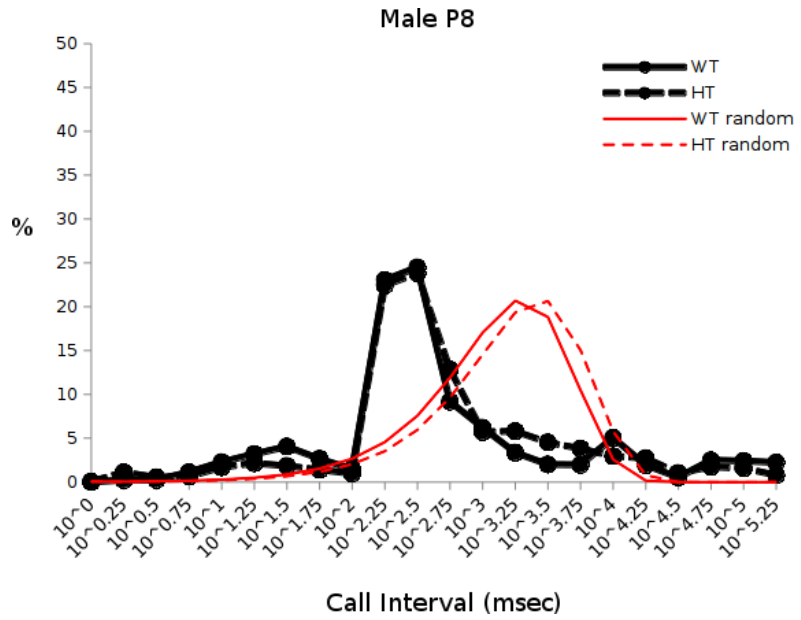

Figure 4.9: Interval Analysis Distributions for randomly generated and original call interval frequencies are shown for the male, postnatal day-8 (MP8) subgroup of mice. The point at which the distributions cross defines the length of the 'gap' between two distinct call strings. (Source: Hiroi Lab)

4.3.2 Unsupervised Analysis using an Information-Theoretic Approach

In order to determine a) if any structure exists within the call strings, b) at what level such structure might be apparent, and c) if there were any discernible differences in structure between the WT and HT phenotypes, we employed a Shannon entropy-based [378] approach. This type of entropy-based

approach to modelling animal vocal calls has previously been applied to mice [379], frogs [380], and in a more exotic fashion, in the search for extraterrestrial life [381]. The entropy of a signal is a measure of its information content, or, in terms of a random variable, a measure of its associated uncertainty. If there is no inherent structure in the vocal call sequences i.e. they are essentially random noise, we would expect the entropy rate to remain flat. If however, there is some level of signal embedded in the calls we should see a deviation from this constant. By assessing the entropy rate using models which incorporate increasing amounts of contextual information about the relative frequencies of the call type configurations possible, we should be able to determine the entropy level at which this structure appears. To that end, we developed a command-line application called Mumbles (manuscript in preparation), which, when given a defined alphabet and a text file of input call strings, will calculate the zeroth to n-th order entropy scores for the call data where n is defined by the user. Here, we use Mumbles to calculate up to fourth-order entropy for mice in each of the eight phenotypic groups with the entropy level cut-off being chosen based on previously published data [379]. In the zeroth-order model, the entropy (in bits) is simply calculated as the *log* of the alphabet size (number of call types, denoted m): $H_0 = log_2 m$. In the first-order model, each of the call types are considered statistically independent and the entropy is based on the single call frequencies:

$$H_1 = -\sum_{i=1}^{m}(p_i)log_2(p_i) \tag{4.1}$$

In the second-order model, conditional probabilities (where $p_{j|i}$ indicates the probability of observing call type j given that call type i has just been emitted) are incorporated to expand the entropy calculation to include two-call, or bi-gram, frequencies. This is equivalent to representing the relationship between two call types as a first-order Markov chain i.e. the probability of an observation x_{n+1} is based only on the probability of observing x_n as the previous state:

$$H_2 = -\sum_{i=1}^{m}(p_i)\sum_{j=1}^{m}(p_{j|i})log_2(p_{j|i}) \tag{4.2}$$

Similary, the third-order entropy model includes a second-order Markov chain component $p_{k|j,i}$, the conditional probability of observing call type k, given that calls j and i preceded it:

$$H_3 = -\sum_{i=1}^{m}(p_i)\sum_{j=1}^{m}(p_{j|i})\sum_{k=1}^{m}(p_{k|j,i})log_2(p_{k|j,i}) \tag{4.3}$$

and so on up to the H_4 level. Once the entropy values for each of the groups have been calculated from H_0 to H_4 (Figure 4.10), we wish to compare the entropy scores at each consecutive order to

determine if there is a statistical decrease when moving from one entropy level to the next.

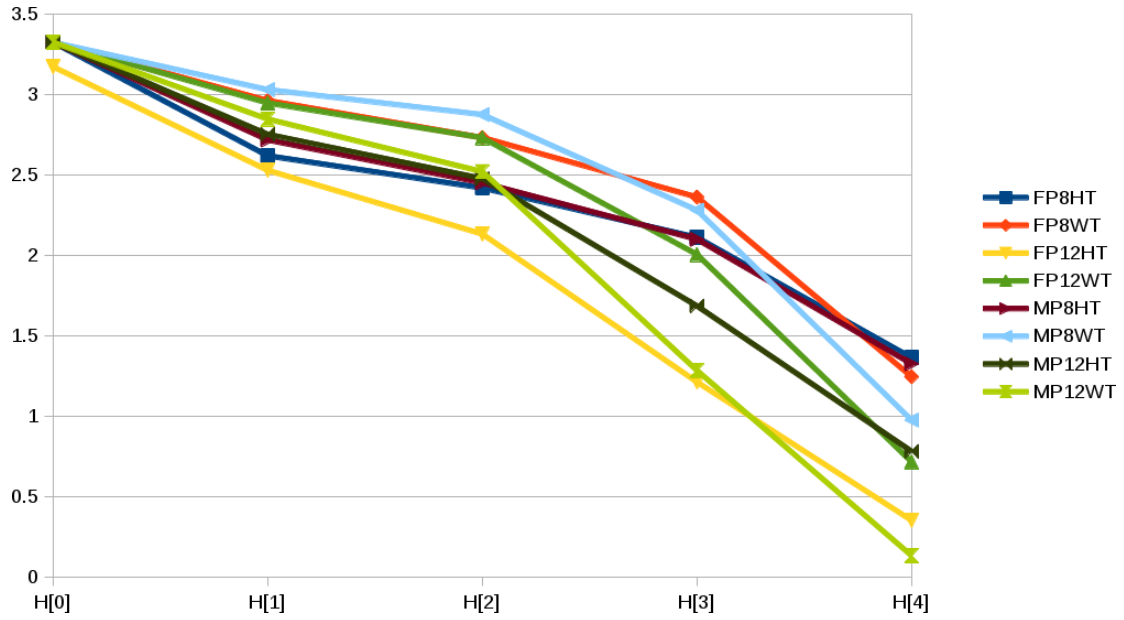

Figure 4.10: Entropy 1 Entropy scores at from H_0 to H_4 are plotted demonstrating deviation from a 'flat' profile and indicating inherent structure in the call sequences.

In order to determine significance, we use the non-parametric two-sample Kolmogorov-Smirnov (K-S) test to ascertain whether or not the entropy scores at two consecutive levels represent samples from the same underlying distribution. Moving from H_0 to H_1, for example, the decrease in entropy has an associated p-value of 0.0006, indicating that the frequencies of the single call types do not occur at the level one would expect based solely on the size of the alphabet. While there is a general decrease in entropy from H_1 to H_2, this decrease is not significant at a p-value cutoff of 0.05 (p=0.0870). The remaining two entropy level comparisons, while significant (p=0.024 for H_2 to H_3 and p=0.0186 H_3 to H_4) also raise an interesting question. From [289], we know that the HT mice have a more restricted vocabulary (use a more limited number of call types) and should therefore show lower entropy scores than their WT counterparts. The relationship evident between HT and WT samples at H_1 and H_2 is consistent with that model; at H_3 and H_4 however, WT mice (particularly in the MP12 and FP12 groups) show much lower entropy scores than expected. Upon further examination, it was discovered that this issue stems from the fact that, as can be seen in Figure 4.11, the majority of the call strings used in the analysis are in fact quite short (< 5 calls per sequence) and cover an extremely limited range of the possible call combinations. The FP8HT group, for example, has the most sequences out of all of the phenotypic groups (838), but for sequences of length four, only 6% of the 10,000 possible call combinations are observed, and of those, more than two thirds appear only once. To examine the effect of this data sparsity, we re-calculated entropy scores to include all possible combinations of call types, including the addition of a pseudocount to

avoid $log(0)$ probabilities. Figure 4.12, which shows the effect of adjusting the psuedocount in the entropy calculations from 1.0 to 0.1, demonstrates a large shift at the H_3 and H_4 levels indicating that sparsity of call data at those levels may point towards a lack of credibility in the statistical determination. The H_2 level entropy scores, while modified by the change in pseudocount, are still statistically drawn from the same distribution (as measured by the K-S test). Taking this combined stability in the face of adjusted pseudocount values and previously indicated trend towards decreased entropy scores (Figure 4.10) we therefore determined H_2 to be the highest entropy order at which structure could reasonably be deemed present in the call sequences without additional data.

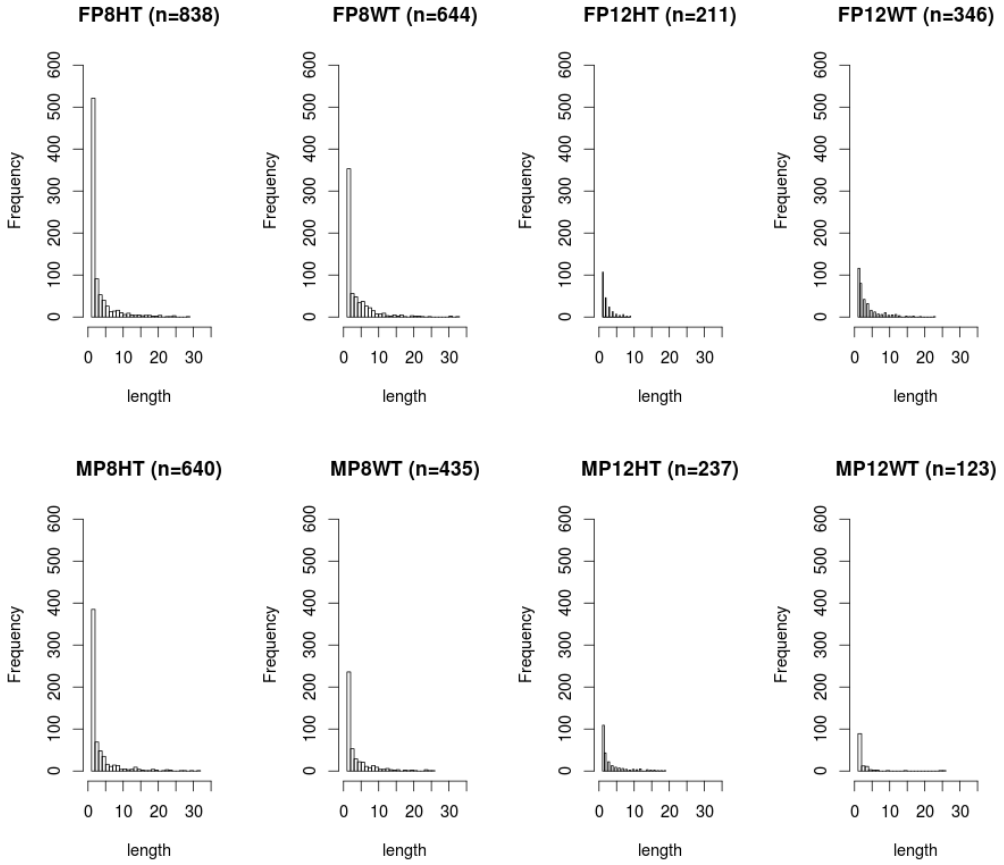

Figure 4.11: Sequence Length Distributions Histograms showing the distributions of sequence length amongst the call strings analyzed indicate that i) there is a great disparity in the number of sequences available for the different phenotype groups, and ii) the majority of the sequences, regardless of phenotypic group, are quite short.

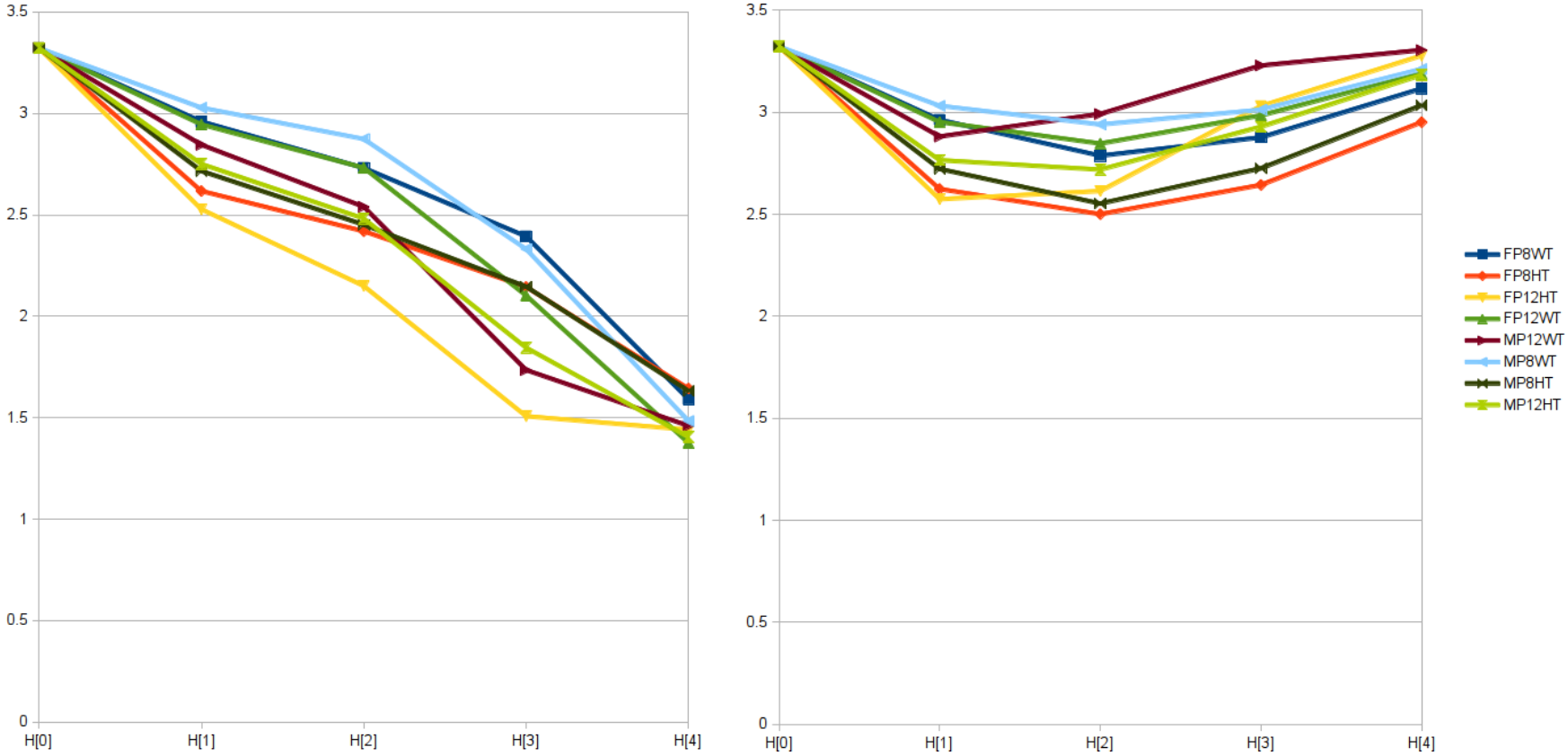

Figure 4.12: Entropy 2 Entropy scores at H_3 and H_4 are disproportionately affected by the addition of a modified pseudocount (left=1.0, right=0.1) indicating an issue with data sparsity.

In order to test if the structure identified in the call sequences is truly necessary for eliciting a maternal response or if the specific frequencies of the different call types are sufficient, 1000 random permutations of an individual WT string were generated. This permutation process maintains the interval and duration associated with each of the individual call types in the original sequence but disrupts the overall structure producing a distribution of entropy values as seen in Figure 4.13. From this distribution three sequences which fall in the higher tail were chosen to be presented in a maternal response test. If the mother demonstrates a response to the perturbed sequences then this will provide evidence that the call frequency alone is sufficient without regard for the higher level sequence structure. At the time of writing, these experiments are currently being carried out by collaborators in Japan.

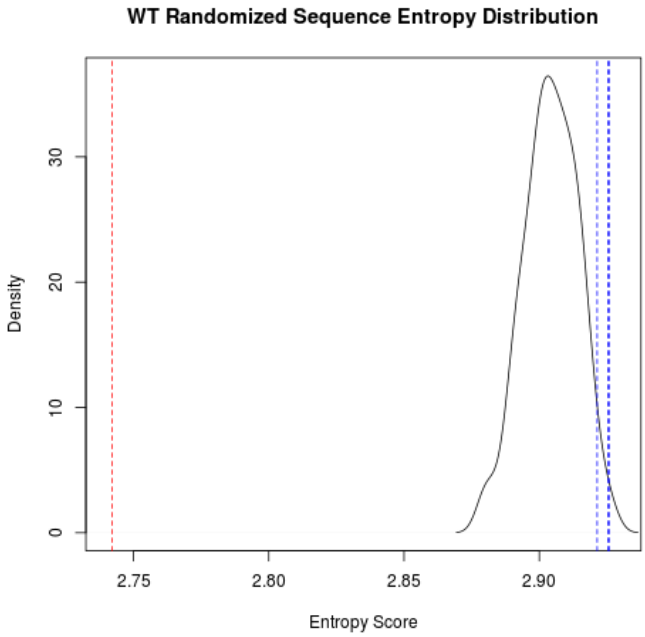

Figure 4.13: Entropy Distribution 1000 permutations of an individual WT sequence results in an entropy distribution from which three sequences (shown in blue) are selected. The sequences selected are those which show the greatest distance from the entropy of the original sequence (shown in red).

4.3.3 Supervised Analysis of Bi-Grams using Projection to Latent Structures

Having identified in an unsupervised manner that the call sequences demonstrate structure up to the second-order level, we then sought to determine in a supervised way which of the bigrams were the most important for distinguishing WT from HT mice. In order to do this we used PLS, a multiple

regression technique which was first introduced by Herman Wold [382] for analysis in the social sciences but which has since become a popular tool in the field of chemometrics. PLS, alternatively known as partial least squares regression or projection to latent structures, is concerned with relating two matrices, X, a set of predictors or independent variables, and Y, a set of dependent or response variables (or, in other words, solving the linear equation $Y = XB + E$, where E represent the residuals). Depending on the dataset in question, this type of analysis may be accomplished using standard multiple linear regression (MLR) and obtaining B from the normal equations ($(X^TX)B = X^TY$), however oftentimes in the 'omics space we are faced with data where a) the number of variables is much greater than the number of observations ($n << p$), and b) multicollinearity exists among the predictor variables [383], in cases such as these the use of MLR is problematic due to the X^TX matrix being ill-conditioned (defined as having a high condition number, or a high ratio of relative change or error in X due to potentially much smaller relative changes in B). A principal components regression (using PCA analysis on X and then performing a regression of Y on the PCs) can alleviate these problems, but, as it only performs a decomposition on X, we can end up with features which explain the variation in X well but are not optimal for prediction of Y. PLS instead combines features of both of these approaches, extracting successive linear combinations of X and Y called latent variables (also known as factors or scores) such that their covariance is maximised. This can be also be thought of as similar to doing a PCA analysis separately on X and Y but including a rotation of the loadings (regression coefficients) in order to maximise their covariance. More specifically, given an $(n \times p)$ matrix X and an $(n \times q)$ matrix Y, these matrices are decomposed as follows:

$$X = TP^T + E \tag{4.4}$$

$$Y = UQ^T + F \tag{4.5}$$

where T and U are matrices of the extracted score vectors or latent components, chosen for maximal covariance, P and Q are loading matrices, and E and F are the matrices of residuals. The process to calculate the decomposition of X and Y can be carried out in a number of ways, one such example is the popular NIPALS (non-linear iterative partial least squares) [382] algorithm which can be summarised as follows: randomly initialise u, the vector of Y-scores, then, while not converged, repeat:

Algorithm 2 NIPALS

1: $w = X^Tu/u^Tu$ ▷ estimate X-weights by regressing onto u
2: $w := w/||w||$ ▷ normalise weights w
3: $t = Xw$ ▷ estimate X-scores
4: $c = Y^Tt/t^Tt$ ▷ estimate Y-weights by regressing onto t
5: $c := c/||c||$ ▷ normalise weights c
6: $u = Yc$ ▷ estimate Y-scores

Once converged (the difference in t between iteration n and $n+1$ is less than some value ϵ), X can be regressed on t to obtaining loadings ($p = X^T t/t^T t$) and b, the regression coefficient used to predict Y from t can be obtained as $b = t^T u/t^T t$. The residuals matrices, E and F, are then calculated as the deflated X and Y matrices based on the current set of latent variables and their associated loadings ($E = X - tp^T$, $F = Y - uq^T$) and the calculated vectors (t, u, w, c, p, and b) are used to populate their respective matrices. This process is then repeated with X and Y being replaced by E and F until as many latent variables as are required have been generated. This can be an efficient approach when dealing with either large datasets where calculating all pairs of latent variables is too time consuming, or in situations where only the first few latent variable pairs are needed to explain the majority of the covariance in a dataset. An alternative to the iterative approach of NIPALS involves the use of singular value decomposition (SVD) to decompose the covariance matrix $X^T Y$ into $U\Sigma V^T$, where Σ is a diagonal matrix of singular values and $U^T U = V^T V = I$. This approach can be particularly beneficial if all factors are required, it is however more computationally costly than NIPALS, which uses iteration to avoid operations directly on the covariance matrix.

While initially proposed for regression analysis, PLS has also been shown to perform well on classification problems [384, 385, 386]. When applied to cases where Y is a vector of categorical variables, this approach is termed partial least squares discriminant analysis, or PLS-DA. A sparse version of PLS is also possible by incorporating an L_1 (or lasso) regularisation when computing the SVD of the covariance matrix. This results in many of the regression coefficients being driven to zero (automatic variable selection) and may lead to improved model interpretability and prediction accuracy. Here, we make use of a combination of these features, applying a sparse PLS-DA (sPLS-DA) approach (as implemented in [387]) to simultaneously perform classification and variable selection on a matrix of bi-gram frequencies, with each row in the frequency matrix X being associated with an entry in the vector of class labels Y. This allows us to identify a subset of vocal calls which provide the best class separation between WT and HT call sequences and to measure the relative importance of these features on class prediction.

We began the analysis by first filtering out calls which show near-zero variance. These are calls which are not likely to be of value as predictors due to their having either only one unique value, or relatively few unique values with a large ratio in the frequency of the most common value to the second most common value. The 17 filtered call types are shown in Table 4.6.

In order to identify both a) the optimal number of components needed to explain the Y, and b) the number of variables to keep in the regression model, (the regularisation parameter λ is automatically chosen by the algorithm based on these selections), we need to perform a cross-validation to assess the overall model error rate for each combination. Figure 4.14 shows the results of a leave-one-out cross validation (LOOCV) which shows the error rate as a function of number of selected calls and number of components (or latent variables) selected. Note that while combinations of up to 10 components and 50 calls (representing half of the possible call combinations) were tested, the error rate only increased beyond the values shown in the plot and so these combinations are not shown.

Call	Ratio	Unique (%)
U-Ha	70	8.00
C-Ts	71	6.67
Ha-U	70	8.00
C-U	0	1.33
Fs-U	71	6.67
C-H	72	5.33
Ts-C	71	6.67
U-C	73	4.00
D-C	69	9.33
Sh-C	70	8.00
Fs-C	70	8.00
F-C	72	5.33
U-Fs	72	5.33
Sh-Fs	69	9.33
C-Fs	71	6.67
F-Fs	69	9.33
C-F	70	8.00

Table 4.6: Filtered Calls Shown here are the 17 calls filtered from the list of 100 call combinations for having near-zero variance. Ratio indicates the ratio of the frequency of the most common value to the frequency of the second most common value while Unique indicates the percentage of samples which have unique values for this call type. U-Ha, for example, has only six different frequency values across the 75 mice, one of these values is zero, which occurs in 70 out of 75 vocal samples. Similarly, the C-U call type is not observed in any of the vocal samples and so has only one unique value, zero.

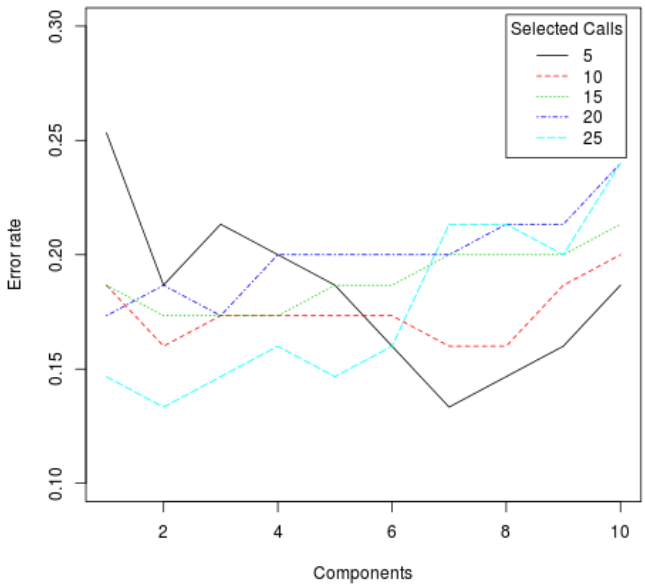

Figure 4.14: Crossvalidation Error rates for the sPLS-DA predictions using a leave-one-out crossvalidation and selecting different combinations of number of call types and latent variables. The minimum error rate of 13.3% was achieved selecting 25 calls across two components.

Once the crossvalidation has been carried out we examine the scores and loadings plots for the sPLS-DA analysis using the selected number of components and call types (Figure 4.15). As in a PCA analysis, the scores plot provides a summary of the relationship between the samples and shows their projection into the new space defined by the indicated pair of latent variables, while the loading plot shows the relationships between the predictor variables and their contribution to this projection. The WT mice, shown in red, are largely separated from their HT counterparts, shown in black, along the first component. They demonstrate a much greater variability in vocal calls than the HT mice, which are tightly clustered in the lower right corner of the plot. This is consistent with the reduced number of call types used by the HT mice.

The loadings plot (Figure 4.16) is shown as a correlation circle, the further a call sequence is along a particular axis, the greater the correlation between that call type and the separation of the samples along that axis as shown in the scores plot. Here, we see that calls such as F-F, Cx-C, C-CX, Cx-F, Fs-F, and Sh-F all have correlation scores greater than 0.5 indicating that they are the most important for separating the WT mice from the HT mice. A further demonstration of this can be seen in the example given in Figure 4.17 which shows a boxplot of the frequencies for the F-F call type in WT and HT mice. We also applied this sPLS-DA approach to the individual gender and age subgroups to determine if any particular group or groups showed any other discernible structure

among the sample types. As shown in Figure 4.18, in each case the HT samples were more tightly clustered than their WT counterparts and the primary separation along the first latent variable axis was between WT and HT samples.

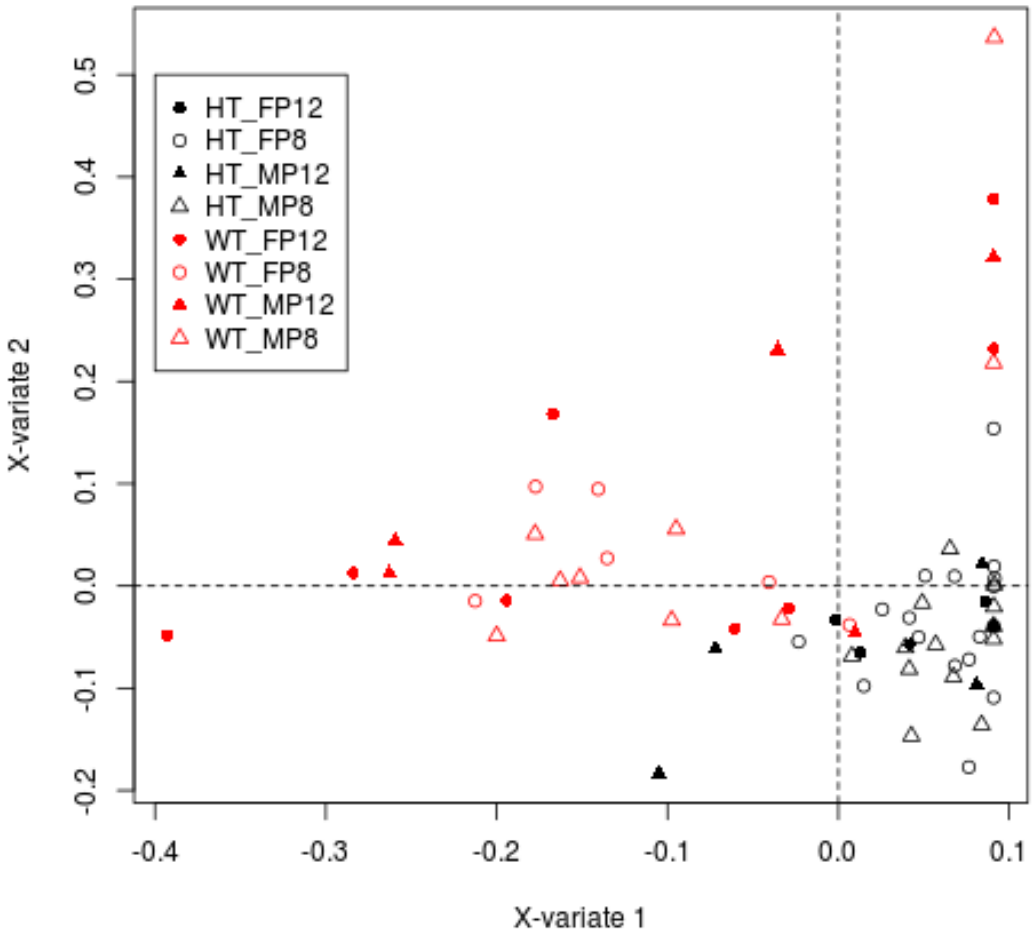

Figure 4.15: sPLS-DA Scores This plot shows the projection of the samples into the space defined by the first two latent variables. WT mice are shown in red and demonstrate much greater vocal variability, HT mice are shown in black.

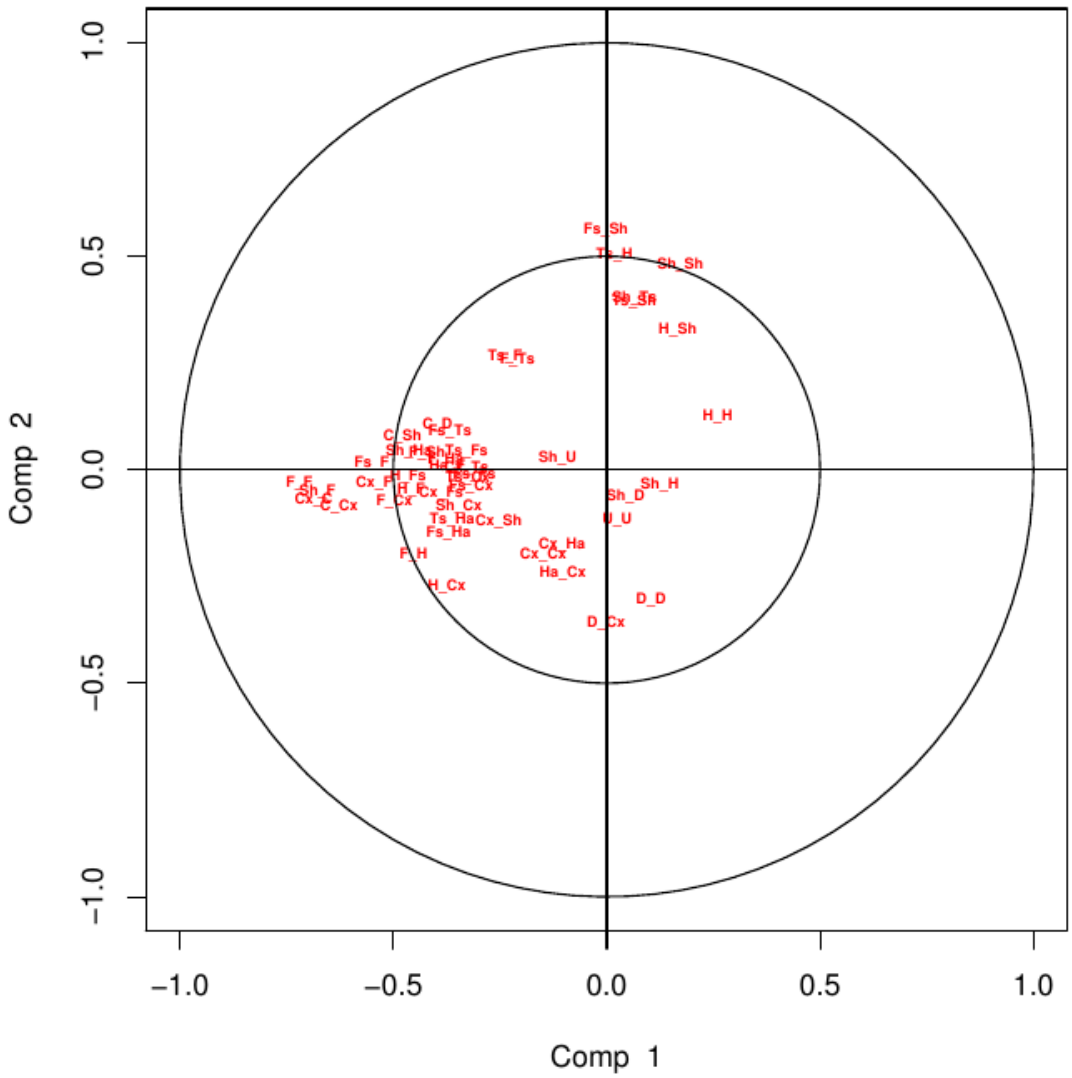

Figure 4.16: sPLS-DA Loadings The correlation circle for the variables (calls) is shown with the distance of each call type along each axis providing an indication of its strength of association with scores lying in that direction.

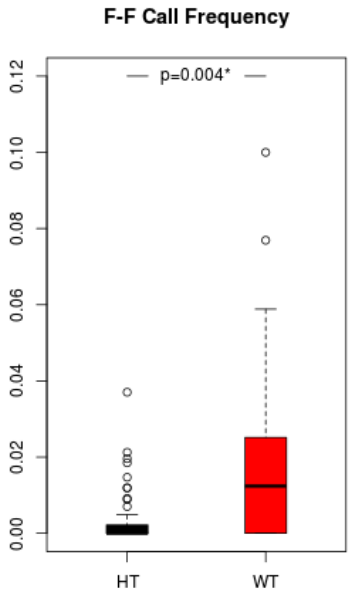

Figure 4.17: Frequency Boxplot A boxplot showing the frequency of F-F call type occurences in both WT and HT samples indicates a clear distinction, with an associated p-value of 0.004.

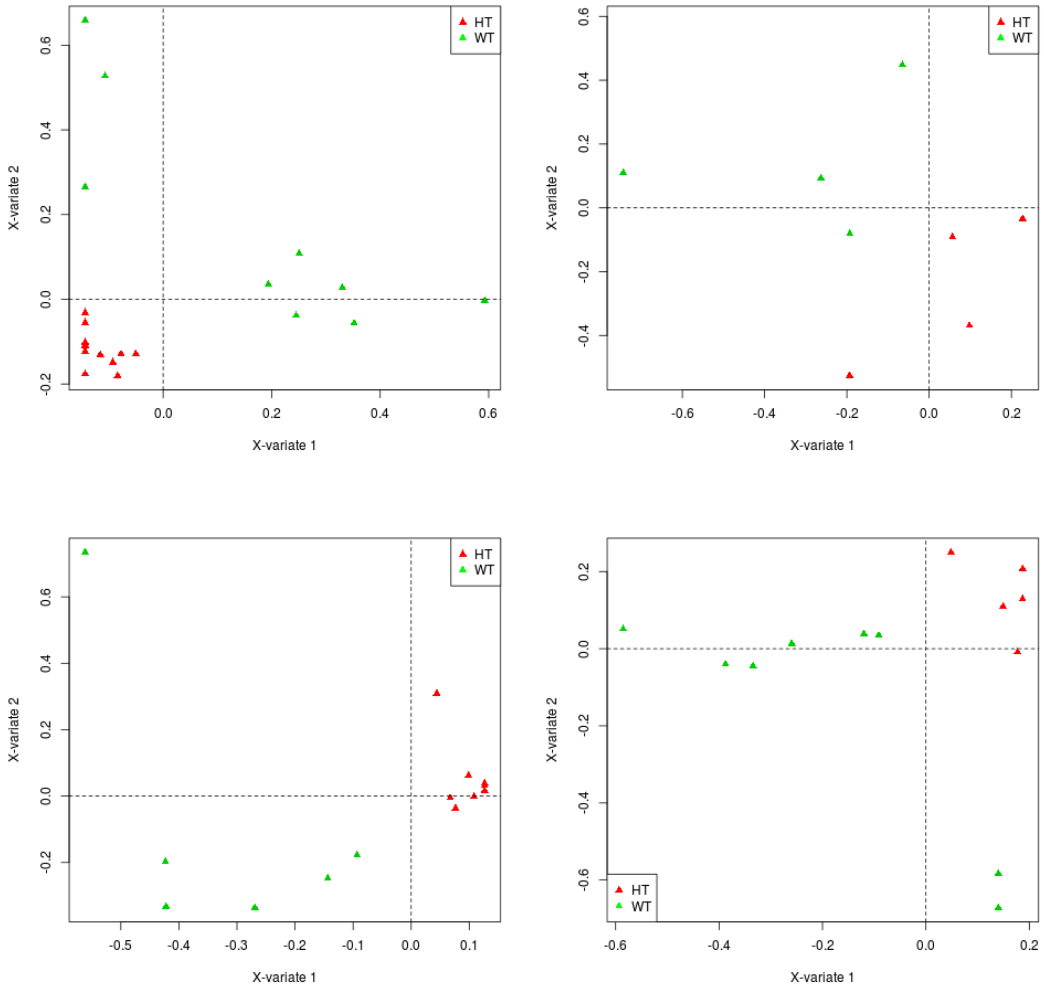

Figure 4.18: Subgroup Analysis The individual subgroups (clockwise from upper left: MP8, MP12, FP8, and FP12) show separation primarily based on phenotype as well as tighter clustering of the HT samples.

4.4 Discussion

22q11DS - A Catch-All Syndrome?

This chapter provided an examination of the links between the genetic role of Tbx1 in postnatal neurogenesis and the social behaviours resulting from Tbx1 haploinsufficiency which mimic several psychiatric disorders including schizophrenia and ASD. A key long-term goal of this study is not only to further our understanding of the role of Tbx1 in these disorders, but also to use any insights gained in that regard to increase our ability to predict clinical phenotypes. As pointed out by Sinderberry et al. [388] however, 22q11.2 deletion syndrome has over 180 associated characteristics, making such predictions difficult. In a recent study of 50 children aged 6-17 years they show however, that two major clinical subtypes with differing degrees of risk for development of psychiatric disorders can be identified. Statistically significant differences between these two subtypes exist in terms of both cognitive features such as IQ, mathematical reasoning, verbal reasoning, and interactive sociability, as well as physical features such temporal cortex and whole brain volume, facial size, and nasal profile. The authors further show a significant difference in terms of presence of autistic traits between these two groups and argue that many of the 22q11DS patients in other studies may have been labelled as ASD based on only a few ASD-like features while not warranting a clinical diagnosis of ASD. This view is further expressed by Angkustsiri et al. [389] who argue that some studies which report diagnoses of ASD in 22q11DS cases, do so solely on the basis of the ADI-R. They argue that the 'gold standard' for ASD diagnosis includes both the ADI-R interview with parents as well as the Autism Diagnostic Observation Schedule, or ADOS, which was first introduced in [390] and further updated in 2000 [391] and 2009 [392] (the latter being designed to allow testing of younger children). The ADOS, administered by a trained psychologist, is designed to allow the tester to observe and evaluate social and communication behaviours in suspected ASD cases in such a way that it is independent of language development. It is also worth pointing out that while the ADOS-2 was released in 2012, it will likely take quite some time before a sufficient number of clinicians are trained in its administration and studies carried out to determine the effects of any change in scoring mechanisms on reported ASD cases in 22q11DS.

Tbx1 ChIP-seq

The WASP system, described in Chapter Two, was used to perform the primary analysis of the Tbx1 ChIP-seq data in postnatal neural progenitor cells. In particular, we draw attention to the quality plots which are automatically generated as part of the sequencing pipeline and were used in this instance to identify a potential problem with the library construction. Despite the successful trimming and re-alignment of the generated reads resulting in a higher total number of MACS-called peaks, we still found that the overall agreement between peak lists among the biological replicates was quite poor. Aside from potential problems during library construction, this issue may also relate to the fact that a commercial ChIP-grade antibody for Tbx1 is not currently available resulting in an inconsistent immunoprecipitation. The recent generation of a Tbx1-GFP fusion protein however

[393] may help to address this issue in future experiments through use of an anti-GFP rather than anti-Tbx1 antibody. Despite these benchwork issues, the Tbx1 targets identified provide a tantalising insight into the interactions of this transcription factor with other key regulators of embryonic growth and development. The genes and CNVs implicated highlight the complicated links, both biological and clinical, between adult neurogenesis, autism, schizophrenia, and related disorders, and underpin the need for a better understanding of the role which Tbx1 may potentially play in their etiology.

While some standard ChIP and qPCR validations were carried out on these potential Tbx1 targets, future plans involve a more thorough genome-wide evaluation of expression changes in Tbx1-deficient mice. Data from an RNA-seq assay could then be integrated with ChIP-seq data resulting in a clearer overall picture of the functional role of Tbx1.

An unexpected result from this analysis was the retrieval of matches to T-box motifs in both the JASPAR and TRANSFAC database for ChIPSOM but not for CudaMEME motif models. While we had assumed that both algorithms would identify T-box motifs, the result is encouraging and suggests that despite MEME being the *de facto* standard for motif discovery, ChIPSOM presents a viable and useful alternative and remains worthy of further development.

Vocal Analysis

The entropy analysis carried out revealed that inherent structure within the call sequences was present up to the second-order (or bigram) level. While structure may indeed be present beyond this level, the sparsity of data available for higher-order entropy calculations makes such a determination problematic. Further analysis on a greater number of longer call sequences will help to resolve this question.

Mumbles, the command-line utility we developed, while providing a simple entropy-based analysis of call sequences has nevertheless proven to be a useful tool and its further development as a web application to allow researchers to easily upload and analyze their data in this manner is currently under way. Such a tool would be equally applicable to studies comparing vocalisation patterns in different strains of mice, as well as for comparing mice at different ages, or in different environmental settings. Alternative approaches to animal vocal analysis, particularly in the case of birdsong, involve the use of Hidden Markov Models [394], although the structure in a birdsong is much more readily identifiable in terms of recurrent motifs and therefore potentially more amenable to this type of analysis.

In terms of algorithms for supervised classification of call sequences, many alternatives are available, including, for example, Random Forest [395] and Support Vector Machines [396]. A sPLS-DA approach was chosen however based on its use in some recent work where we have had success in applying it to metabolic profiles in blood plasma in order to determine a dose-predictive biomarker signature for radiation induced damage in whole body gamma-irradiated mice (manuscript submitted to International Journal of Radiation Oncology*Biology*Physics).

CHAPTER 5

Conclusion

Summary & Future Directions

This thesis has provided an overview of the informatics challenges associated with the analysis of next-generation sequencing data for the genome-wide study of genetic and epigenetic regulatory mechanisms. In Chapter One we discussed some of these mechanisms and described their importance in all aspects of biological functioning, as well as the essential role which knowledge of them will play in the era of personalised medicine. This understanding is currently being expanded by large scale and consortia-based projects such as The 1000 Genomes, ENCODE, modENCODE, and IHEC projects (amongst others), which are leading the way to uncovering how these mechanisms differ between organisms and between individuals, in development and ageing, and in health and disease. One of the most challenging aspects of these studies relates to the fact that while each individual has a single genome, they possess many epigenomes. Epigenetic differences not only in different cell types, but also in response to different environmental stimuli results in a highly dynamic system of transcriptional regulation, the deciphering of which will require not only an enormous amount of benchwork but also massive compute and storage resources as well. This will involve reliance on fast processing storage in the form of solid state drives (SSDs) as well as the increasing use of specialised hardware such as field programmable gate arrays (FPGAs) and heterogeneous computing technologies including Nvidia GP-GPUs and co-processor architectures like Intel's Xeon Phi.

Chapter Two detailed WASP, which provides an end-to-end informatics infrastructure to support the automated processing and primary analysis of NGS data. Using ChIP-seq as an example, we demonstrated the various aspects of the system, including sample submission and tracking, core facility LIMS functionality, and backend processing pipelines. The current phase of this project, The WASP System and The WASP Swarm, include the redevelopment of the core platform as an extensible Spring-based plugin architecture and its roll-out to several partner institutions for testing. Once this test phase has been completed, both the core system and plugin API will be made publicly

available for download and the process of community development will begin. The configurable nature of this next evolution of the system means that it can be readily used and expanded not only for new genomics and epigenomics assays, but also for proteomics, metabolomics, and other high-throughput work. The current challenge therefore, lies not in the automated primary analysis of high-throughput biological data, but in the integration of multiple, disparate datasets to create a systems level understanding of biological processes and to generate new testable hypotheses. This integration and mining of multiple large datasets will require not only scalable algorithms, but will also necessitate the development of novel visualisation tools to allow bench researchers to interact with their data in a more direct and intuitive way. One such example of this is the use of space-filling curves to transform the 1D representation of locus based information as used in most genome browsers to a 2D image format which lends itself to standard image processing and manipulation techniques.

In Chapter Three we proposed the new ChIPSOM and GMACS algorithms for the secondary analysis of data from CHIP-seq experiments. We have previously demonstrated the successful use of ChIPSOM on ChIP-chip dataset for the WT1 and CTCF binding factors, and in Chapter Four have demonstrated its use and favourable comparison to MEME on a ChIP-seq dataset for the genome-wide binding of Tbx1. Despite its success however, the algorithm remains computationally costly and future work will seek to improve the runtime either through further code optimisation or the re-development of the code for use on specialised hardware. Future development will also likely move from the SOM implementation to the more flexible growing neural gas (GNG) approach which avoids the problems associated with a fixed node structure and size. Aside from motif discovery in ChIP-datasets, we have also recently used the ChIPSOM approach for pattern discovery in RNA aptamer studies where SELEX [397], or systematic evolution of ligands by exponential enrichment, has been combined with sequencing to determine RNA structures necessary for high-affinity binding of target ligands.

While initially envisaged solely as part of the ChIPSOM algorithm, GMACS has proven useful in its own right for the discovery of relationships between structurally similar classes of transcription factors. We have shown it to be competitive with current field-leading algorithms and will further develop it as a web application to make it more widely available to the research community. The k-mer frequency vector (KFV) metric for sequence comparison has also proven effective and combining it with a GNG trained on sequences from known bacterial species represents a potentially interesting approach to metagenomic analysis which we are currently investigating.

Chapter Four introduced 22q11.2 Deletion Syndrome and outlined its genetic basis as well as the importance of the Tbx1 transcription factor in its etiology. We provided results from one of the first genome-wide studies of Tbx1 binding in postnatal neural progenitor cells and demonstrate the multiple links the potentially targeted genes show to neural development, adult neurogenesis, ASD, and schizophrenia. As indicated in that chapter, future work will involve confirmation of further high-confidence target genes as well as a full transcriptional profiling using RNA-seq to allow us to link Tbx1 binding to changes in gene expression. The vocal analysis section described the use of entropy-based analysis and supervised classification to both demonstrate inherent structure and

differences in the vocal emissions of P8 and P12 WT and HT mouse pups and to enable the potential phenotyping of mice based on call sequences. The tool we developed, Mumbles, will also be deployed as a web application and its use is planned for further studies in the analysis of vocal patterns from different mouse strains. A simple Hidden Markov Model approach is also currently being investigated which would include a gap or 'silence' state, obviating the need to determine sequence cutpoints based on theoretical distributions.

In summary, the increasing adoption of genome-wide sequencing technologies is both incredibly exciting and incredibly challenging, bringing us tantalisingly close to the reality of medical treatments specifically tailored to an individual's genetic and epigenetic makeup, while also pushing the limits of current computational approaches. As the use of these technologies continues to expand our fundamental knowledge of human health and disease, a corresponding fundamental re-imagining of how we store, process, interpret, and interact with the associated data will also be required, driving advances in data mining, machine learning, and visualisation and bringing with it new understanding from amidst great complexity.

Looking Forward - Moving Towards a Systems Approach

Within the broader fields of computer science, neuroscience, genetics and epigenetics, this work lies primarily at the intersection of machine learning and HPC and relates to their use specifically in the analysis of next-generation sequencing data and its application to the study of regulatory mechanisms at a genome scale. While this use is in itself a promising step forward in understanding the global effects and function of a particular transcription factor, histone modification, or chromatin conformation, even taking all of this information to the next logical stage of data integration will still present only a relatively limited and static view of what is, at its core, an inherently dynamic system. With researchers coming to appreciate the limitations of such a static approach, focus has begun to shift in recent years to computational systems biology in which attempts are made to model complex and dynamic interactions between large numbers of individual system components. One example of this is the increasing utilisation of gene regulatory and protein-protein interaction (PPI) networks to provide a more holistic view of regulation than is possible when considering only one individual gene or protein at a time, even when considered at a genome-scale. Using information from individual experiments to both build and infer relationships between genes and proteins has resulted in rich regulatory networks which can be mined to provide further biological insights through, for example, graph-based analysis resulting in the identification of hubs (which may function as master regulators), sub-networks, or redundant interaction pathways (which may point to more robust regulatory mechanisms necessary to maintain key biological processes). These networks can also serve to generate hypotheses for bench validation by predicting regulatory outcomes and changes in expression based on perturbations to single or multiple nodes within the networks – this is a key aspect of virtual screening for therapeutic agents which can result in both marked savings for pharmaceutical companies as well as greatly decreased time to identification of promising lead compounds.

This type of systems approach has also been used to create biologically realistic models of both individual cells as well as multi-cellular systems. CytoSolve[1] was one of the first attempts at a large-scale integrative model to incorporate a wide range of the complex interactions in a whole human cell, including metabolomic, transcriptomic, and gene regulatory networks. It makes use of systems biology markup language (SBML) and has been tested and successfully applied to multiple in silico drug discovery projects [398]. Similarly, a whole cell model of the bacterium Mycoplasma genitalium was created in 2012 which includes all 525 annotated genes and their interactions, encompassing 28 biological processes and allowing the simulation of the entire life cycle of the bacterium, providing new insights into previously unobserved cellular behaviours [399]. Naturally, complex models of this nature rely heavily on HPC resources, and this particular model, when run on a 128-cores, takes approximately 10 hours to simulate a single cell division. The OpenWorm[2] community is currently pursuing a similar project which aims to produce an accurate in silico model of all 959 cells in C. elegans.

Neuroscience, and in particular research focusing on the human brain with its vast number of neurons and myriad inter-connections is another prime example of where this type of systems modelling is helping to uncover new insights. The ambitious Human Brain Project[3] with a budget of almost $1.5 billion aims to construct a full working model of the human brain within the next 10 years. A key challenge this project faces is the inability of current hardware to sufficiently mimic the asynchronous nature of signalling which occurs in spiking neurons. Digital signals are by design binary in nature (on/off, high/low); one of the Human Brain Project's goals therefore is to advance research is what is termed neuromorphic computing, with the aim of creating integrated circuits containing electronic analog components (such as neuristors) which more accurately reflect the non-linear activation of neurons in the biological nervous system. Running simulations for such a large model would also require inordinate compute resources, in the order of exascale computing. Researchers at the RIKEN institute in Japan for example, home to the fourth most powerful supercomputer in the world, containing over 700,000 cores and 1.4 million GB RAM, recently managed to simulate one second of brain activity using a model of approximately 1.7 billion neurons and 10.4 trillion synapses. This sub-model equates to roughly 1% of the total model required to capture the complexity of the human brain and took more than 40 minutes to run on 82,944 processor cores[4]. Given such vast technological requirements, it remains to be seen if such a lofty goal is truly feasible within this predicted time frame, and if so, how such a model might be realistically used to relate fundamental signalling processes in distinct brain regions to complex cognitive behaviour.

[1] http://www.cytosolve.com/
[2] http://www.openworm.org/
[3] http://www.humanbrainproject.eu/
[4] http://www.riken.jp/

Bibliography

[1] R. Staden. A strategy of dna sequencing employing computer programs. *Nucleic Acids Res*, 6(7):2601–2610, Jun 1979.

[2] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269(5223):496–512, Jul 1995.

[3] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guig, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291(5507):1304–1351, Feb 2001.

[4] J. D. McPherson, M. Marra, L. Hillier, R. H. Waterston, A. Chinwalla, J. Wallis, M. Sekhon, K. Wylie, E. R. Mardis, R. K. Wilson, R. Fulton, T. A. Kucaba, C. Wagner-McPherson, W. B. Barbazuk, S. G. Gregory, S. J. Humphray, L. French, R. S. Evans, G. Bethel, A. Whittaker, J. L. Holden, O. T. McCann, A. Dunham, C. Soderlund, C. E. Scott, D. R. Bentley, G. Schuler, H. C. Chen, W. Jang, E. D. Green, J. R. Idol, V. V. Maduro, K. T. Montgomery, E. Lee, A. Miller, S. Emerling, Kucherlapati, R. Gibbs, S. Scherer, J. H. Gorrell, E. Sodergren, K. Clerc-Blankenburg, P. Tabor, S. Naylor, D. Garcia, P. J. de Jong, J. J. Catanese, N. Nowak, K. Osoegawa, S. Qin, L. Rowen, A. Madan, M. Dors, L. Hood, B. Trask, C. Friedman, H. Massa, V. G. Cheung, I. R. Kirsch, T. Reid, R. Yonescu, J. Weissenbach, T. Bruls, R. Heilig, E. Branscomb, A. Olsen, N. Doggett, J. F. Cheng, T. Hawkins, R. M. Myers, J. Shang, L. Ramirez, J. Schmutz, O. Velasquez, K. Dixon, N. E. Stone, D. R. Cox, D. Haussler, W. J. Kent, T. Furey, S. Rogic, S. Kennedy, S. Jones, A. Rosenthal, G. Wen, M. Schilhabel, G. Gloeckner, G. Nyakatura, R. Siebert, B. Schlegelberger, J. Korenberg, X. N. Chen, A. Fujiyama, M. Hattori,

A. Toyoda, T. Yada, H. S. Park, Y. Sakaki, N. Shimizu, S. Asakawa, K. Kawasaki, T. Sasaki, A. Shintani, A. Shimizu, K. Shibuya, J. Kudoh, S. Minoshima, J. Ramser, P. Seranski, C. Hoff, A. Poustka, R. Reinhardt, H. Lehrach, and International Human Genome Mapping Consortium. A physical map of the human genome. *Nature*, 409(6822):934–941, Feb 2001.

[5] Ioanna Pagani, Konstantinos Liolios, Jakob Jansson, I-Min A Chen, Tatyana Smirnova, Bahador Nosrat, Victor M Markowitz, and Nikos C Kyrpides. The genomes online database (gold) v. 4: status of genomic and metagenomic projects and their associated metadata. *Nucleic acids research*, 40(D1):D571–D579, 2012.

[6] Michael L. Metzker. Sequencing technologies the next generation. *Nat Rev Genetics*, 11:31–46, 2010.

[7] Malin Akerfelt, Eva Henriksson, Asta Laiho, Anniina Vihervaara, Karoliina Rautoma, Noora Kotaja, and Lea Sistonen. Promoter chip-chip analysis in mouse testis reveals y chromosome occupancy by hsf2. *Proc Natl Acad Sci U S A*, 105(32):11224–11229, Aug 2008.

[8] J. D. Lieb, X. Liu, D. Botstein, and P. O. Brown. Promoter-specific binding of rap1 revealed by genome-wide maps of protein-dna association. *Nat Genet*, 28(4):327–334, Aug 2001.

[9] Dirk Schubeler, David M MacAlpine, David Scalzo, Christiane Wirbelauer, Charles Kooperberg, Fred van Leeuwen, Daniel E Gottschling, Laura P O'Neill, Bryan M Turner, Jeffrey Delrow, Stephen P Bell, and Mark Groudine. The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev*, 18(11):1263–1271, Jun 2004.

[10] Stefan Tmpel, Leanne M Wiedemann, and Robb Krumlauf. Hox genes and segmentation of the vertebrate hindbrain. *Curr Top Dev Biol*, 88:103–137, 2009.

[11] K. Niederreither, J. Vermot, B. Schuhbaur, P. Chambon, and P. Doll. Retinoic acid synthesis and hindbrain patterning in the mouse embryo. *Development*, 127(1):75–85, Jan 2000.

[12] Q. Zhou, G. Choi, and D. J. Anderson. The bhlh transcription factor olig2 promotes oligodendrocyte differentiation in collaboration with nkx2.2. *Neuron*, 31(5):791–807, Sep 2001.

[13] A. M. Reimold, N. N. Iwakoshi, J. Manis, P. Vallabhajosyula, E. Szomolanyi-Tsuda, E. M. Gravallese, D. Friend, M. J. Grusby, F. Alt, and L. H. Glimcher. Plasma cell differentiation requires the transcription factor xbp-1. *Nature*, 412(6844):300–307, Jul 2001.

[14] Tadanobu Nagaya, Naoki Tanaka, Takefumi Suzuki, Kenji Sano, Akira Horiuchi, Michiharu Komatsu, Takero Nakajima, Tomoko Nishizawa, Satoru Joshita, Takeji Umemura, Tetsuya Ichijo, Akihiro Matsumoto, Kaname Yoshizawa, Jun Nakayama, Eiji Tanaka, and Toshifumi Aoyama. Down-regulation of srebp-1c is associated with the development of burned-out nash. *J Hepatol*, 53(4):724–731, Oct 2010.

[15] Koichi Yagi, Kiwamu Akagi, Hiroshi Hayashi, Genta Nagae, Shingo Tsuji, Takayuki Isagawa, Yutaka Midorikawa, Yoji Nishimura, Hirohiko Sakamoto, Yasuyuki Seto, Hiroyuki Aburatani, and Atsushi Kaneda. Three dna methylation epigenotypes in human colorectal cancer. *Clin Cancer Res*, 16(1):21–33, Jan 2010.

[16] Paul E Neiman, Katrina Elsaesser, Gilbert Loring, and Robert Kimmel. Myc oncogene-induced genomic instability: Dna palindromes in bursal lymphomagenesis. *PLoS Genet*, 4(7):e1000132, 2008.

[17] Mario F Fraga and Manel Esteller. Epigenetics and aging: the targets and the marks. *Trends Genet*, 23(8):413–418, Aug 2007.

[18] J. David Sweatt. Neuroscience. epigenetics and cognitive aging. *Science*, 328(5979):701–702, May 2010.

[19] Chunhui Hou, Hui Zhao, Keiji Tanimoto, and Ann Dean. Ctcf-dependent enhancer-blocking by alternative chromatin loop formation. *Proceedings of the National Academy of Sciences*, 105(51):20398–20403, 2008.

[20] R. Tjian. Molecular machines that control genes. *Sci Am*, 272(2):54–61, Feb 1995.

[21] F. H. Crick. On protein synthesis. *Symp Soc Exp Biol*, 12:138–163, 1958.

[22] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, Aug 1970.

[23] M. Madan Babu, Nicholas M Luscombe, L. Aravind, Mark Gerstein, and Sarah A Teichmann. Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol*, 14(3):283–291, Jun 2004.

[24] Priyadarshi Basu, Thanh Giang Sargent, Latasha C Redmond, Jeremy C Aisenberg, Evan P Kransdorf, Shou Zhen Wang, Gordon D Ginder, and Joyce A Lloyd. Evolutionary conservation of klf transcription factors and functional conservation of human gamma-globin gene regulation in chicken. *Genomics*, 84(2):311–319, Aug 2004.

[25] G. D. Amoutzias, A. S. Veron, J. Weiner, M. Robinson-Rechavi, E. Bornberg-Bauer, S. G. Oliver, and D. L. Robertson. One billion years of bzip transcription factor evolution: conservation and change in dimerization and dna-binding site specificity. *Mol Biol Evol*, 24(3):827–835, Mar 2007.

[26] Erik van Nimwegen. Scaling laws in the functional content of genomes. *Trends Genet*, 19(9):479–484, Sep 2003.

[27] Benjamin P Berman, Yutaka Nibu, Barret D Pfeiffer, Pavel Tomancak, Susan E Celniker, Michael Levine, Gerald M Rubin, and Michael B Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the drosophila genome. *Proc Natl Acad Sci U S A*, 99(2):757–762, Jan 2002.

[28] G. Gill. Regulation of the initiation of eukaryotic transcription. *Essays Biochem*, 37:33–43, 2001.

[29] S. J. Busch and P. Sassone-Corsi. Dimers, leucine zippers and dna-binding domains. *Trends Genet*, 6(2):36–40, Feb 1990.

[30] Derek Jantz, Barbara T Amann, Gregory J Gatto, and Jeremy M Berg. The design of functional dna-binding proteins based on zinc finger domains. *Chem Rev*, 104(2):789–799, Feb 2004.

[31] Geeta J Narlikar, Hua-Ying Fan, and Robert E Kingston. Cooperation between complexes that regulate chromatin structure and transcription. *Cell*, 108(4):475–487, Feb 2002.

[32] Yeon Hee Kang, Victor Kirik, Martin Hulskamp, Kyoung Hee Nam, Katherine Hagely, Myeong Min Lee, and John Schiefelbein. The myb23 gene provides a positive feedback loop for cell fate specification in the arabidopsis root epidermis. *Plant Cell*, 21(4):1080–1094, Apr 2009.

[33] Barbara Lustig, Boris Jerchow, Martin Sachs, Sigrid Weiler, Torsten Pietsch, Uwe Karsten, Marc van de Wetering, Hans Clevers, Peter M Schlag, Walter Birchmeier, and Jrgen Behrens. Negative feedback loop of wnt signaling through upregulation of conductin/axin2 in colorectal and liver tumors. *Mol Cell Biol*, 22(4):1184–1193, Feb 2002.

[34] H. E. Huber, G. Edwards, P. J. Goodhart, D. R. Patrick, P. S. Huang, M. Ivey-Hoyle, S. F. Barnett, A. Oliff, and D. C. Heimbrook. Transcription factor e2f binds dna as a heterodimer. *Proc Natl Acad Sci U S A*, 90(8):3525–3529, Apr 1993.

[35] Klaus H Hansen, Adrian P Bracken, Diego Pasini, Nikolaj Dietrich, Simmi S Gehani, Astrid Monrad, Juri Rappsilber, Mads Lerdrup, and Kristian Helin. A model for transmission of the h3k27me3 epigenetic mark. *Nat Cell Biol*, 10(11):1291–1300, Nov 2008.

[36] Raphael Margueron, Neil Justin, Katsuhito Ohno, Miriam L Sharpe, Jinsook Son, William J Drury, Philipp Voigt, Stephen R Martin, William R Taylor, Valeria De Marco, Vincenzo Pirrotta, Danny Reinberg, and Steven J Gamblin. Role of the polycomb protein eed in the propagation of repressive histone marks. *Nature*, 461(7265):762–767, Oct 2009.

[37] Eva Jablonka and Gal Raz. Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. *Q Rev Biol*, 84(2):131–176, Jun 2009.

[38] K. Luger, A. W. Mder, R. K. Richmond, D. F. Sargent, and T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8 a resolution. *Nature*, 389(6648):251–260, Sep 1997.

[39] Tae-Young Roh and Keji Zhao. High-resolution, genome-wide mapping of chromatin modifications by gmat. *Methods Mol Biol*, 387:95–108, 2008.

[40] Tae-Young Roh, Suresh Cuddapah, Kairong Cui, and Keji Zhao. The genomic landscape of histone modifications in human t cells. *Proc Natl Acad Sci U S A*, 103(43):15782–15787, Oct 2006.

[41] B. D. Strahl and C. D. Allis. The language of covalent histone modifications. *Nature*, 403(6765):41–45, Jan 2000.

[42] T. Jenuwein and C. D. Allis. Translating the histone code. *Science*, 293(5532):1074–1080, Aug 2001.

[43] Elizaveta V Benevolenskaya. Histone h3k4 demethylases are essential in development and differentiation. *Biochem Cell Biol*, 85(4):435–443, Aug 2007.

[44] Christoph M Koch, Robert M Andrews, Paul Flicek, Shane C Dillon, Ula Karaz, Gayle K Clelland, Sarah Wilcox, David M Beare, Joanna C Fowler, Phillippe Couttet, Keith D James, Gregory C Lefebvre, Alexander W Bruce, Oliver M Dovey, Peter D Ellis, Pawandeep Dhami, Cordelia F Langford, Zhiping Weng, Ewan Birney, Nigel P Carter, David Vetrie, and Ian Dunham. The landscape of histone modifications across 1in five human cell lines. *Genome Res*, 17(6):691–707, Jun 2007.

[45] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, May 2007.

[46] Jeffrey A Rosenfeld, Zhibin Wang, Dustin E Schones, Keji Zhao, Rob DeSalle, and Michael Q Zhang. Determination of enriched histone modifications in non-genic portions of the human genome. *BMC Genomics*, 10:143, 2009.

[47] Cassandra Hogan and Patrick Varga-Weisz. The regulation of atp-dependent nucleosome remodelling factors. *Mutat Res*, 618(1-2):41–51, May 2007.

[48] Karl P Nightingale, Matthias Baumann, Anton Eberharter, Adamantios Mamais, Peter B Becker, and Joan Boyes. Acetylation increases access of remodelling complexes to their nucleosome targets to enhance initiation of v(d)j recombination. *Nucleic Acids Res*, 35(18):6311–6321, 2007.

[49] Anton Eberharter, Roger Ferreira, and Peter Becker. Dynamic chromatin: concerted nucleosome remodelling and acetylation. *Biol Chem*, 386(8):745–751, Aug 2005.

[50] Steven A Jacobs and Sepideh Khorasanizadeh. Structure of hp1 chromodomain bound to a lysine 9-methylated histone h3 tail. *Science*, 295(5562):2080–2083, Mar 2002.

[51] D. O. Jones, I. G. Cowell, and P. B. Singh. Mammalian chromodomain proteins: their role in genome organisation and expression. *Bioessays*, 22(2):124–137, Feb 2000.

[52] E. Li, C. Beard, and R. Jaenisch. Role for dna methylation in genomic imprinting. *Nature*, 366(6453):362–365, Nov 1993.

[53] T. Goto and M. Monk. Regulation of x-chromosome inactivation in development in mice and humans. *Microbiol Mol Biol Rev*, 62(2):362–378, Jun 1998.

[54] E. Scarano, M. Iaccarino, P. Grippo, and E. Parisi. The heterogeneity of thymine methyl group origin in dna pyrimidine isostichs of developing sea urchin embryos. *Proc Natl Acad Sci U S A*, 57(5):1394–1400, May 1967.

[55] A. P. Bird. Dna methylation and the frequency of cpg in animal dna. *Nucleic Acids Res*, 8(7):1499–1504, Apr 1980.

[56] M. Ehrlich, M. A. Gama-Sosa, L. H. Huang, R. M. Midgett, K. C. Kuo, R. A. McCune, and C. Gehrke. Amount and distribution of 5-methylcytosine in human dna from different types of tissues of cells. *Nucleic Acids Res*, 10(8):2709–2721, Apr 1982.

[57] Mehrnaz Fatemi, Martha M Pao, Shinwu Jeong, Einav Nili Gal-Yam, Gerda Egger, Daniel J Weisenberger, and Peter A Jones. Footprinting of mammalian promoters: use of a cpg dna methyltransferase revealing nucleosome positions at a single molecule level. *Nucleic Acids Res*, 33(20):e176, 2005.

[58] Daiya Takai and Peter A Jones. Comprehensive analysis of cpg islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A*, 99(6):3740–3745, Mar 2002.

[59] P. L. Jones, G. J. Veenstra, P. A. Wade, D. Vermaak, S. U. Kass, N. Landsberger, J. Strouboulis, and A. P. Wolffe. Methylated dna and mecp2 recruit histone deacetylase to repress transcription. *Nat Genet*, 19(2):187–191, Jun 1998.

[60] P. W. Laird and R. Jaenisch. The role of dna methylation in cancer genetic and epigenetics. *Annu Rev Genet*, 30:441–464, 1996.

[61] Toshinori Hinoue, Daniel J Weisenberger, Fei Pan, Mihaela Campan, Myungjin Kim, Joanne Young, Vicki L Whitehall, Barbara A Leggett, and Peter W Laird. Analysis of the association between cimp and braf in colorectal cancer by dna methylation profiling. *PLoS One*, 4(12):e8357, 2009.

[62] P. A. Jones and P. W. Laird. Cancer epigenetics comes of age. *Nat Genet*, 21(2):163–167, Feb 1999.

[63] Hiroaki Kawasaki and Kazunari Taira. Microrna-196 inhibits hoxb8 expression in myeloid differentiation of hl60 cells. *Nucleic Acids Symp Ser (Oxf)*, (48):211–212, 2004.

[64] A. E. Williams. Functional aspects of animal micrornas. *Cell Mol Life Sci*, 65(4):545–562, Feb 2008.

[65] Ana Eulalio, Eric Huntzinger, Tadashi Nishihara, Jan Rehwinkel, Maria Fauser, and Elisa Izaurralde. Deadenylation is a widespread effect of mirna regulation. *RNA*, 15(1):21–32, Jan 2009.

[66] R. C. Lee, R. L. Feinbaum, and V. Ambros. The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *Cell*, 75(5):843–854, Dec 1993.

[67] B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun. The 21-nucleotide let-7 rna regulates developmental timing in caenorhabditis elegans. *Nature*, 403(6772):901–906, Feb 2000.

[68] Anna M Krichevsky, Kevin S King, Christine P Donahue, Konstantin Khrapko, and Kenneth S Kosik. A microrna array reveals extensive regulation of micrornas during brain development. *RNA*, 9(10):1274–1281, Oct 2003.

[69] Jose Dostie, Zissimos Mourelatos, Michael Yang, Anup Sharma, and Gideon Dreyfuss. Numerous micrornps in neuronal cells containing novel micrornas. *RNA*, 9(2):180–186, Feb 2003.

[70] Xuemei Chen. A microrna as a translational repressor of apetala2 in arabidopsis flower development. *Science*, 303(5666):2022–2025, Mar 2004.

[71] Julius Brennecke, David R Hipfner, Alexander Stark, Robert B Russell, and Stephen M Cohen. bantam encodes a developmentally regulated microrna that controls cell proliferation and regulates the proapoptotic gene hid in drosophila. *Cell*, 113(1):25–36, Apr 2003.

[72] Peizhang Xu, Stephanie Y Vernooy, Ming Guo, and Bruce A Hay. The drosophila microrna mir-14 suppresses cell death and is required for normal fat metabolism. *Curr Biol*, 13(9):790–795, Apr 2003.

[73] Lin He, J. Michael Thomson, Michael T Hemann, Eva Hernando-Monge, David Mu, Summer Goodson, Scott Powers, Carlos Cordon-Cardo, Scott W Lowe, Gregory J Hannon, and Scott M Hammond. A microrna polycistron as a potential human oncogene. *Nature*, 435(7043):828–833, Jun 2005.

[74] Marek Mraz, Sarka Pospisilova, Karla Malinova, Ivo Slapak, and Jiri Mayer. Micrornas in chronic lymphocytic leukemia pathogenesis and disease subtypes. *Leuk Lymphoma*, 50(3):506–509, Mar 2009.

[75] Markus Metzler, Monika Wilda, Kerstin Busch, Susanne Viehmann, and Arndt Borkhardt. High expression of precursor microrna-155/bic rna in children with burkitt lymphoma. *Genes Chromosomes Cancer*, 39(2):167–169, Feb 2004.

[76] Thomas Thum, Paolo Galuppo, Christian Wolf, Jan Fiedler, Susanne Kneitz, Linda W van Laake, Pieter A Doevendans, Christine L Mummery, Jrgen Borlak, Axel Haverich, Carina Gross, Stefan Engelhardt, Georg Ertl, and Johann Bauersachs. Micrornas in the human heart: a clue to fetal gene reprogramming in heart failure. *Circulation*, 116(3):258–267, Jul 2007.

[77] Alessandra Car, Daniele Catalucci, Federica Felicetti, Dsire Bonci, Antonio Addario, Paolo Gallo, Marie-Louise Bang, Patrizia Segnalini, Yusu Gu, Nancy D Dalton, Leonardo Elia, Michael V G Latronico, Morten Hydal, Camillo Autore, Matteo A Russo, Gerald W Dorn, Oyvind Ellingsen, Pilar Ruiz-Lozano, Kirk L Peterson, Carlo M Croce, Cesare Peschle, and Gianluigi Condorelli. Microrna-133 controls cardiac hypertrophy. *Nat Med*, 13(5):613–618, May 2007.

[78] Olivier C Maes, Howard M Chertkow, Eugenia Wang, and Hyman M Schipper. Microrna: Implications for alzheimer disease and other human cns disorders. *Curr Genomics*, 10(3):154–168, May 2009.

[79] N. J. Beveridge, E. Gardiner, A. P. Carroll, P. A. Tooney, and M. J. Cairns. Schizophrenia is associated with an increase in cortical microrna biogenesis. *Mol Psychiatry*, Sep 2009.

[80] Andrea Tanzer and Peter F Stadler. Molecular evolution of a microrna cluster. *J Mol Biol*, 339(2):327–335, May 2004.

[81] Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, and David P Bartel. Most mammalian mrnas are conserved targets of micrornas. *Genome Res*, 19(1):92–105, Jan 2009.

[82] Pasquale Fasanaro, Simona Greco, Mircea Ivan, Maurizio C Capogrossi, and Fabio Martelli. microrna: emerging therapeutic targets in acute ischemic diseases. *Pharmacol Ther*, 125(1):92–104, Jan 2010.

[83] N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel. An abundant class of tiny rnas with probable regulatory roles in caenorhabditis elegans. *Science*, 294(5543):858–862, Oct 2001.

[84] Antony Rodriguez, Sam Griffiths-Jones, Jennifer L Ashurst, and Allan Bradley. Identification of mammalian microrna host genes and transcription units. *Genome Res*, 14(10A):1902–1910, Oct 2004.

[85] Yoontae Lee, Chiyoung Ahn, Jinju Han, Hyounjeong Choi, Jaekwang Kim, Jeongbin Yim, Junho Lee, Patrick Provost, Olof Rdmark, Sunyoung Kim, and V. Narry Kim. The nuclear rnase iii drosha initiates microrna processing. *Nature*, 425(6956):415–419, Sep 2003.

[86] E. Lund and J. E. Dahlberg. Substrate selectivity of exportin 5 and dicer in the biogenesis of micrornas. *Cold Spring Harb Symp Quant Biol*, 71:59–66, 2006.

[87] K. Okamura, M.D. Phillips, D.M. Tyler, H. Duan, and Chou Y.T. et al. The regulatory activity of microrna star species has substantial influence on microrna and 3 ' utr evolution. *Nature Structural & Molecular Biology*, 15:354363, 2008.

[88] Dianne S Schwarz, Gyrgy Hutvgner, Tingting Du, Zuoshang Xu, Neil Aronin, and Phillip D Zamore. Asymmetry in the assembly of the rnai enzyme complex. *Cell*, 115(2):199–208, Oct 2003.

[89] Shaun Mahony, David Hendrix, Aaron Golden, Terry J Smith, and Daniel S Rokhsar. Transcription factor binding site identification using the self-organizing map. *Bioinformatics*, 21(9):1807–1814, May 2005.

[90] T. Kohonen. *Self-organizing maps,3rd ed.* Berlin, New York:Springer, 2001.

[91] David E Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning.* Boston, MA:Kluwer Academic Publishers, 1989.

[92] A. S Fraser. Simulation of genetic systems by automatic digital computers. i. introduction. *Aust J Biol Sci*, 10:484–491, 1957.

[93] W.-L. Tam and B. Lim. *StemBook*, chapter Genome-wide transcription factor localization and function in stem cells. 2008. doi/10.3824/stembook.1.19.1.

[94] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, Jun 2007.

[95] Peter J Park. Chip-seq: advantages and challenges of a maturing technology. *Nat Rev Genetics*, 10(10):669–680, Oct 2009.

[96] S. Fields. Molecular biology. site-seeing by sequencing. *Science*, 316(5830):1441–2, 2007.

[97] Christopher A Maher, Nallasivam Palanisamy, John C Brenner, Xuhong Cao, Shanker Kalyana-Sundaram, Shujun Luo, Irina Khrebtukova, Terrence R Barrette, Catherine Grasso, Jindan Yu, Robert J Lonigro, Gary Schroth, Chandan Kumar-Sinha, and Arul M Chinnaiyan. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc Natl Acad Sci U S A*, 106(30):12353–12358, Jul 2009.

[98] Thomas A Down, Vardhman K Rakyan, Daniel J Turner, Paul Flicek, Heng Li, Eugene Kulesha, Stefan Grf, Nathan Johnson, Javier Herrero, Eleni M Tomazou, Natalie P Thorne, Liselotte Bckdahl, Marlis Herberth, Kevin L Howe, David K Jackson, Marcos M Miretti, John C Marioni, Ewan Birney, Tim J P Hubbard, Richard Durbin, Simon Tavar, and Stephan Beck. A bayesian deconvolution strategy for immunoprecipitation-based dna methylome analysis. *Nat Biotechnol*, 26(7):779–785, Jul 2008.

[99] M. Frommer, L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy, and C. L. Paul. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual dna strands. *Proc Natl Acad Sci U S A*, 89(5):1827–1831, Mar 1992.

[100] Masako Suzuki, Qiang Jing, Daniel Lia, Marin Pascual, Andrew McLellan, and John M Greally. Optimized design and data analysis of tag-based cytosine methylation assays. *Genome Biol*, 11(4):R36, 2010.

[101] Heng Li, Jue Ruan, and Richard Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–1858, Nov 2008.

[102] B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome Res*, 8(3):186–194, Mar 1998.

[103] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009.

[104] Stephen M Rumble, Phil Lacroute, Adrian V Dalca, Marc Fiume, Arend Sidow, and Michael Brudno. Shrimp: accurate mapping of short color-space reads. *PLoS Comput Biol*, 5(5):e1000386, May 2009.

[105] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197, Mar 1981.

[106] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[107] M. Burrows and D.J. Wheeler. A block sorting lossless data compression algorithm. Technical report, Digital Equipment Corporation, 1994.

[108] Heng Li and Richard Durbin. Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*, 26(5):589–595, Mar 2010.

[109] Michael C Schatz. Cloudburst: highly sensitive read mapping with mapreduce. *Bioinformatics*, 25(11):1363–1369, Jun 2009.

[110] Andrew D Smith, Zhenyu Xuan, and Michael Q Zhang. Using quality scores and longer reads improves accuracy of solexa read mapping. *BMC Bioinformatics*, 9:128, 2008.

[111] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[112] Teemu D Laajala, Sunil Raghav, Soile Tuomela, Riitta Lahesmaa, Tero Aittokallio, and Laura L Elo. A practical comparison of methods for detecting transcription factor binding sites in chip-seq experiments. *BMC Genomics*, 10:618, 2009.

[113] Elizabeth G Wilbanks and Marc T Facciotti. Evaluation of algorithm performance in chip-seq peak detection. *PLoS One*, 5(7):e11471, 2010.

[114] Joel Rozowsky, Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B Gerstein. Peakseq enables systematic scoring of chip-seq experiments relative to controls. *Nat Biotechnol*, 27(1):66–75, Jan 2009.

[115] Raja Jothi, Suresh Cuddapah, Artem Barski, Kairong Cui, and Keji Zhao. Genome-wide identification of in vivo protein-dna binding sites from chip-seq data. *Nucleic Acids Res*, 36(16):5221–5231, Sep 2008.

[116] Yong Zhang, Tao Liu, Clifford A Meyer, Jrme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X. Shirley Liu. Model-based analysis of chip-seq (macs). *Genome Biol*, 9(9):R137, 2008.

[117] Hongkai Ji, Hui Jiang, Wenxiu Ma, David S Johnson, Richard M Myers, and Wing H Wong. An integrated software system for analyzing chip-chip and chip-seq data. *Nat Biotechnol*, 26(11):1293–1300, Nov 2008.

[118] Anton Valouev, David S Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M Myers, and Arend Sidow. Genome-wide analysis of transcription factor binding sites based on chip-seq data. *Nat Methods*, 5(9):829–834, Sep 2008.

[119] Zhaohui S Qin, Jianjun Yu, Jincheng Shen, Christopher A Maher, Ming Hu, Shanker Kalyana-Sundaram, Jindan Yu, and Arul M Chinnaiyan. Hpeak: an hmm-based algorithm for defining read-enriched regions in chip-seq data. *BMC Bioinformatics*, 11:369, 2010.

[120] Andrew S McLellan, Robert A Dubin, Qiang Jing, Pilib Ó Broin, David Moskowitz, Masako Suzuki, R Brent Calder, Joseph Hargitai, Aaron Golden, and John M Greally. The wasp system: An open source environment for managing and analyzing genomic data. *Genomics*, 2012.

[121] Andrea Sboner, Xinmeng Jasmine Mu, Dov Greenbaum, Raymond K Auerbach, Mark B Gerstein, et al. The real cost of sequencing: higher than you think. *Genome Biol*, 12(8):125, 2011.

[122] Brent Ewing, LaDeana Hillier, Michael C Wendl, and Phil Green. Base-calling of automated sequencer traces usingphred. i. accuracy assessment. *Genome research*, 8(3):175–185, 1998.

[123] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[124] Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F Kim, Alexandra Soboleva, Maxim Tomashevsky, Kimberly A Marshall, et al. Ncbi geo: archive for high-throughput functional genomic data. *Nucleic acids research*, 37(suppl 1):D885–D890, 2009.

[125] Yuichi Kodama, Martin Shumway, and Rasko Leinonen. The sequence read archive: explosive growth of sequencing data. *Nucleic acids research*, 40(D1):D54–D56, 2012.

[126] W. James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at ucsc. *Genome Res*, 12(6):996–1006, Jun 2002.

[127] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative genomics viewer. *Nature biotechnology*, 29(1):24–26, 2011.

[128] Cole Trapnell, Lior Pachter, and Steven L Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[129] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.

[130] Duncan Hull, Katy Wolstencroft, Robert Stevens, Carole Goble, Mathew R Pocock, Peter Li, and Tom Oinn. Taverna: a tool for building and running workflows of services. *Nucleic acids research*, 34(suppl 2):W729–W732, 2006.

[131] Burkhard Linke, Robert Giegerich, and Alexander Goesmann. Conveyor: a workflow engine for bioinformatic analyses. *Bioinformatics*, 27(7):903–911, 2011.

[132] Aris Floratos, Kenneth Smith, Zhou Ji, John Watkinson, and Andrea Califano. geworkbench: an open source platform for integrative genomics. *Bioinformatics*, 26(14):1779–1780, 2010.

[133] Bernd Jagla, Bernd Wiswedel, and Jean-Yves Coppée. Extending knime for next-generation sequencing data analysis. *Bioinformatics*, 27(20):2907–2909, 2011.

[134] Jeremy Goecks, Anton Nekrutenko, James Taylor, T Galaxy Team, et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8):R86, 2010.

[135] S Gesing et al. The einstein genome gateway using wasp-a high throughput multi-layered life sciences portal for xsede. *HealthGrid Applications and Technologies Meet Science Gateways for Life Sciences*, 175:182, 2012.

[136] David Rhee, Joseph Hargitai, R Brent Calder, Pilib Ó Broin, Kevin Shieh, and Aaron Golden. 'spring through the gateway': deploying genomic workflows with xsede. In *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery*, page 30. ACM, 2013.

[137] Patrik D'haeseleer. What are dna sequence motifs? *Nat Biotech*, 24:423–425, 2006.

[138] Gary B Fogel, Dana G Weekes, Gabor Varga, Ernst R Dow, Andrew M Craven, Harry B Harlow, Eric W Su, Jude E Onyia, and Chen Su. A statistical analysis of the transfac database. *Biosystems*, 81(2):137–154, Aug 2005.

[139] Alan M Moses, Derek Y Chiang, Manolis Kellis, Eric S Lander, and Michael B Eisen. Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol Biol*, 3:19, Aug 2003.

[140] G. A. Babbitt. Relaxed selection against accidental binding of transcription factors with conserved chromatin contexts. *Gene*, 466(1-2):43–48, Oct 2010.

[141] Zeba Wunderlich and Leonid A Mirny. Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet*, 25(10):434–440, Oct 2009.

[142] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, Oct 1993.

[143] Martha L Bulyk, Philip L F Johnson, and George M Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res*, 30(5):1255–1261, Mar 2002.

[144] K. M. Scully, E. M. Jacobson, K. Jepsen, V. Lunyak, H. Viadiu, C. Carrire, D. W. Rose, F. Hooshmand, A. K. Aggarwal, and M. G. Rosenfeld. Allosteric effects of pit-1 dna sites on long-term repression in cell type specification. *Science*, 290(5494):1127–1131, Nov 2000.

[145] Robert Osada, Elena Zaslavsky, and Mona Singh. Comparative analysis of methods for representing and searching for transcription factor binding sites. *Bioinformatics*, 20(18):3516–3525, Dec 2004.

[146] Xiaoyue Zhao, Haiyan Huang, and Terence P Speed. Finding short dna motifs using permuted markov models. *J Comput Biol*, 12(6):894–906, 2005.

[147] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097–6100, Oct 1990.

[148] G. D. Stormo. Dna binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, Jan 2000.

[149] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender. Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue):D108–D110, Jan 2006.

[150] Dominique Vlieghe, Albin Sandelin, Pieter J De Bleser, Kris Vleminckx, Wyeth W Wasserman, Frans van Roy, and Boris Lenhard. A new generation of jaspar, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res*, 34(Database issue):D95–D97, Jan 2006.

[151] Richard Mnch, Karsten Hiller, Heiko Barg, Dana Heldt, Simone Linz, Edgar Wingender, and Dieter Jahn. Prodoric: prokaryotic database of gene regulation. *Nucleic Acids Res*, 31(1):266–269, Jan 2003.

[152] J. Zhu and M. Q. Zhang. Scpd: a promoter database of the yeast saccharomyces cerevisiae. *Bioinformatics*, 15(7-8):607–611, 1999.

[153] Heladia Salgado, Socorro Gama-Castro, Martn Peralta-Gil, Edgar Daz-Peredo, Fabiola Snchez-Solano, Alberto Santos-Zavaleta, Irma Martnez-Flores, Vernica Jimnez-Jacinto, Csar Bonavides-Martnez, Juan Segura-Salazar, Agustino Martnez-Antonio, and Julio Collado-Vides. Regulondb (version 5.0): Escherichia coli k-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res*, 34(Database issue):D394–D397, Jan 2006.

[154] K. Quandt, K. Frech, H. Karas, E. Wingender, and T. Werner. Matind and matinspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res*, 23(23):4878–4884, Dec 1995.

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

[156] Marko Djordjevic, Anirvan M Sengupta, and Boris I Shraiman. A biophysical approach to transcription factor binding site discovery. *Genome Res*, 13(11):2381–2390, Nov 2003.

[157] S. Sinha and M. Tompa. A statistical method for finding transcription factor binding sites. *Proc Int Conf Intell Syst Mol Biol*, 8:344–354, 2000.

[158] T. J. Wu, Y. C. Hsieh, and L. A. Li. Statistical measures of dna sequence dissimilarity under markov chain models of base composition. *Biometrics*, 57(2):441–448, Jun 2001.

[159] Giulio Pavesi, Giancarlo Mauri, and Graziano Pesole. In silico representation and discovery of transcription factor binding sites. *Brief Bioinform*, 5(3):217–236, Sep 2004.

[160] G. Pavesi, G. Mauri, and G. Pesole. An algorithm for finding signals of unknown length in dna sequences. *Bioinformatics*, 17 Suppl 1:S207–S214, 2001.

[161] H. J. Bussemaker, H. Li, and E. D. Siggia. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc Natl Acad Sci U S A*, 97(18):10096–10100, Aug 2000.

[162] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38, 1977.

[163] Timothy L Bailey, Nadya Williams, Chris Misleh, and Wilfred W Li. Meme: discovering and analyzing dna and protein sequence motifs. *Nucleic Acids Res*, 34(Web Server issue):W369–W373, Jul 2006.

[164] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6 (6):721–741, 1984.

[165] F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church. Finding dna regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation. *Nat Biotechnol*, 16(10):939–945, Oct 1998.

[166] Matthieu Defrance and Jacques van Helden. info-gibbs: a motif discovery algorithm that directly optimizes information content during sampling. *Bioinformatics*, 25(20):2715–2722, Oct 2009.

[167] William Thompson, Eric C Rouchka, and Charles E Lawrence. Gibbs recursive sampler: finding transcription factor binding sites. *Nucleic Acids Res*, 31(13):3580–3585, Jul 2003.

[168] X. Shirley Liu, Douglas L Brutlag, and Jun S Liu. An algorithm for finding protein-dna binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol*, 20(8):835–839, Aug 2002.

[169] Albin Sandelin, Wyeth W Wasserman, and Boris Lenhard. Consite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res*, 32(Web Server issue):W249–W252, Jul 2004.

[170] Gabriela G Loots and Ivan Ovcharenko. rvista 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res*, 32(Web Server issue):W217–W221, Jul 2004.

[171] Stein Aerts, Gert Thijs, Bert Coessens, Mik Staes, Yves Moreau, and Bart De Moor. Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res*, 31(6):1753–1764, Mar 2003.

[172] A. M. McGuire, J. D. Hughes, and G. M. Church. Conservation of dna regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res*, 10(6):744–757, Jun 2000.

[173] Emmanouil T Dermitzakis and Andrew G Clark. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol*, 19(7):1114–1121, Jul 2002.

[174] Kurtis Eisermann, Sunpreet Tandon, Anton Bazarov, Adina Brett, Gail Fraizer, and Helen Piontkivska. Evolutionary conservation of zinc finger transcription factor binding sites in promoters of genes co-expressed with wt1 in prostate cancer. *BMC Genomics*, 9:337, 2008.

[175] Anthony R Borneman, Tara A Gianoulis, Zhengdong D Zhang, Haiyuan Yu, Joel Rozowsky, Michael R Seringhaus, Lu Yong Wang, Mark Gerstein, and Michael Snyder. Divergence of transcription factor binding sites across related yeast species. *Science*, 317(5839):815–819, Aug 2007.

[176] Qing Zhou and Wing H Wong. Cismodule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci U S A*, 101(33):12114–12119, Aug 2004.

[177] Nikolaus Rajewsky, Massimo Vergassola, Ulrike Gaul, and Eric D Siggia. Computational detection of genomic cis-regulatory modules applied to body patterning in the early drosophila embryo. *BMC Bioinformatics*, 3:30, Oct 2002.

[178] Saurabh Sinha, Erik van Nimwegen, and Eric D Siggia. A probabilistic method to detect regulatory modules. *Bioinformatics*, 19 Suppl 1:i292–i301, 2003.

[179] Katherine Belov, Janine E Deakin, Anthony T Papenfuss, Michelle L Baker, Sandra D Melman, Hannah V Siddle, Nicolas Gouin, David L Goode, Tobias J Sargeant, Mark D Robinson, Matthew J Wakefield, Shaun Mahony, Joseph G R Cross, Panayiotis V Benos, Paul B Samollow, Terence P Speed, Jennifer A Marshall Graves, and Robert D Miller. Reconstructing an ancestral mammalian immune supercomplex from a marsupial major histocompatibility complex. *PLoS Biol*, 4(3):e46, Mar 2006.

[180] W. W. Wasserman and J. W. Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol*, 278(1):167–181, Apr 1998.

[181] Stein Aerts, Peter Van Loo, Gert Thijs, Yves Moreau, and Bart De Moor. Computational detection of cis -regulatory modules. *Bioinformatics*, 19 Suppl 2:ii5–i14, Oct 2003.

[182] D. Benjamin Gordon, Lena Nekludova, Scott McCallum, and Ernest Fraenkel. Tamo: a flexible, object-oriented framework for analyzing transcriptional regulation using dna-sequence motifs. *Bioinformatics*, 21(14):3164–3165, Jul 2005.

[183] Katherine A Romer, Guy-Richard Kayombya, and Ernest Fraenkel. Webmotifs: automated discovery, filtering and scoring of dna sequence motifs using multiple programs and bayesian approaches. *Nucleic Acids Res*, 35(Web Server issue):W217–W220, Jul 2007.

[184] T. Kohonen. Automatic formation of topological maps of patterns in a self-organizing system. In *2nd Scandinavian Conf. Image Analysis, Espoo, Finland, 15–17 June*, pages 214–220. Helsinki: Pattern Recognition Society of Finland., 1981.

[185] S. Kaski, J. Kangas, and T. Kohonen. Bibliography of self-organizing map (som) papers 1981–1997. *Neural Computing Surveys*, 1:1–176, 1998.

[186] Merja Oja, Samuel Kaski, and Teuvo Kohonen. Bibliography of self-organizing map (som) papers: 1998-2001 addendum. *Neural Computing Surveys*, 3:1–156, 2003.

[187] S Kaski. Computationally efficient approximation of a probabilistic model for document representation in the websom full-text analysis method. *Neural Processing Letters*, 5(2):139–151, 1997.

[188] C. Amerijckx, M. Verleysen, P. Thissen, and J. D. Legat. Image compression by self-organized kohonen map. *IEEE Transactions on Neural Networks*, 9(3):503–507, 1998.

[189] S. Grossberg and J. R. Williamson. A self organizing neural system for learning to recognize textured scenes. *Vision Research*, 39:1385–1406, 1999.

[190] Y. Liu and R. H. Weisberg. Patterns of ocean current variability on the west florida shelf using the self-organizing map. *Journal of Geophysical Research*, 110, 2005.

[191] S. Kaski, J. Sinkkonen, and J. Peltonen. Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks*, 12(4):936–947, 2001.

[192] A. M. Jennings and J. Graham. A neural network approach to automatic chromosome classification. *Physics in Medicine and Biology*, 38(7):959–970, 1993.

[193] H. Ritter. Self-organizing maps on non-euclidean spaces. In E. Oja and S. Kaski, editors, *Kohonen Maps*. Amsterdam:Elsevier, 1999.

[194] T. Kohonen and P. Somervuo. How to make large self-organizing maps for nonvectorial data. *Neural Networks*, 15(8-9):945–952, 2002.

[195] A. Ultsch and H. P. Siemon. Kohonen's self organizing feature maps for exploratory data analysis. In *Proc. Intern. Neural Networks*, pages 305–308, 1990.

[196] Shaun Mahony, Panayiotis V Benos, Terry J Smith, and Aaron Golden. Self-organizing neural networks to support the discovery of dna-binding motifs. *Neural Networks*, 19(6-7):950–962, 2006.

[197] Brandon J. Thomas, Eric D. Rubio, Niklas Krumm, Pilib O. Broin, Karol Bomsztyk, Piri Welcsh, John M. Greally, Aaron A. Golden, and Anton Krumm. Allele-specific transcriptional elongation regulates monoallelic expression of the igf2bp1 gene. *Epigenetics Chromatin*, 4:14, 2011.

[198] Gene M. Amdahl. Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the April 18-20, 1967, Spring Joint Computer Conference*, AFIPS '67 (Spring), pages 483–485, New York, NY, USA, 1967. ACM.

[199] Marianne K-H Kim, Thomas J McGarry, Pilib O Broin, Jared M Flatow, Aaron A-J Golden, and Jonathan D Licht. An integrated genome screen identifies the wnt signaling pathway as a major target of wt1. *Proc Natl Acad Sci U S A*, 106(27):11154–11159, Jul 2009.

[200] Albin Sandelin and Wyeth W Wasserman. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *Journal of molecular biology*, 338(2):207–215, 2004.

[201] Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W Wasserman, and Boris Lenhard. Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, 32(suppl 1):D91–D94, 2004.

[202] Shaun Mahony, Aaron Golden, Terry J Smith, and Panayiotis V Benos. Improved detection of dna motifs using a self-organized clustering of familial binding profiles. *Bioinformatics*, 21 Suppl 1:i283–i291, Jun 2005.

[203] Eric P Xing and Richard M Karp. Motifprototyper: a bayesian profile model for motif families. *Proceedings of the National Academy of Sciences of the United States of America*, 101(29):10523–10528, 2004.

[204] Leelavati Narlikar, Raluca Gordân, Uwe Ohler, and Alexander J Hartemink. Informative priors based on transcription factor structural class improve de novo motif discovery. *Bioinformatics*, 22(14):e384–e392, 2006.

[205] Szymon M Kielbasa, Didier Gonze, and Hanspeter Herzel. Measuring similarities between transcription factor binding sites. *BMC bioinformatics*, 6(1):237, 2005.

[206] Dustin E Schones, Pavel Sumazin, and Michael Q Zhang. Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, 21(3):307–313, 2005.

[207] Shaun Mahony, Philip E Auron, and Panayiotis V Benos. Dna familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS computational biology*, 3(3):e61, 2007.

[208] Utz J Pape, Sven Rahmann, and Martin Vingron. Natural similarity measures between position frequency matrices with an application to clustering. *Bioinformatics*, 24(3):350–357, 2008.

[209] Minli Xu and Zhengchang Su. A novel alignment-free method for comparing transcription factor binding site motifs. *PloS one*, 5(1):e8797, 2010.

[210] Shaun Mahony and Panayiotis V Benos. Stamp: a web tool for exploring dna-binding motif similarities. *Nucleic acids research*, 35(suppl 2):W253–W258, 2007.

[211] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

[212] Robert R Sokal and Charles D Michener. *A statistical method for evaluating systematic relationships*. University of Kansas, 1958.

[213] Geoffrey J Barton and Michael JE Sternberg. A strategy for the rapid multiple alignment of protein sequences: Confidence levels from tertiary structure comparisons. *Journal of molecular biology*, 198(2):327–337, 1987.

[214] Stuart Lloyd. Least squares quantization in pcm. *Information Theory, IEEE Transactions on*, 28(2):129–137, 1982.

[215] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. Wiley. com, 1990.

[216] John H Holland. *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence.* U Michigan Press, 1975.

[217] David E Goldberg and John H Holland. Genetic algorithms and machine learning. *Machine learning*, 3(2):95–99, 1988.

[218] Cédric Notredame and Desmond G Higgins. Saga: sequence alignment by genetic algorithm. *Nucleic acids research*, 24(8):1515–1524, 1996.

[219] Cédric Notredame, Emmet A O'Brien, and Desmond G Higgins. Raga: Rna sequence alignment by genetic algorithm. *Nucleic Acids Research*, 25(22):4570–4580, 1997.

[220] Zhi Wei and Shane T Jensen. Game: detecting cis-regulatory elements using a genetic algorithm. *Bioinformatics*, 22(13):1577–1584, 2006.

[221] Falcon FM Liu, Jeffrey JP Tsai, Rong-Ming Chen, SN Chen, and SH Shih. Fmga: finding motifs by genetic algorithm. In *Bioinformatics and Bioengineering, 2004. BIBE 2004. Proceedings. Fourth IEEE Symposium on*, pages 459–466. IEEE, 2004.

[222] Carlos B Lucasius, Adrie D Dane, and Gerrit Kateman. On¡ i¿ k¡/i¿-medoid clustering of large data sets with the aid of a genetic algorithm: background, feasiblity and comparison. *Analytica Chimica Acta*, 282(3):647–669, 1993.

[223] Weiguo Sheng and Xiaohui Liu. A genetic k-medoids clustering algorithm. *Journal of Heuristics*, 12(6):447–466, 2006.

[224] Jon Bohlin, Eystein Skjerve, and David W Ussery. Investigations of oligonucleotide usage variance within and between prokaryotes. *PLoS computational biology*, 4(4):e1000057, 2008.

[225] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[226] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[227] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[228] Vea Matys, Ellen Fricke, R Geffers, Ellen Gößling, Martin Haubrock, R Hehl, Klaus Hornischer, Dagmar Karas, Alexander E. Kel, Olga V. Kel-Margoulis, et al. Transfac®: transcriptional regulation, from patterns to profiles. *Nucleic acids research*, 31(1):374–378, 2003.

[229] Leelavati Narlikar and Alexander J Hartemink. Sequence features of dna binding sites reveal structural class of associated transcription factor. *Bioinformatics*, 22(2):157–163, 2006.

[230] R. Miikkulainen. Script recognition with hierarchical feature maps. *Connection Science*, 2:83–101, 1990.

[231] H. Chen, C. Schuffels, and R. Orwig. Internet categorization and search: a self-organizing approach. *Journal of Visual Communication and Image Representation*, 7:88–102, 1996.

[232] Bernd Fritzke. Growing grid - a self-organizing network with constant neighborhood range and adaptation strength. *Neural Processing Letters*, 2(5):9–13, 1995.

[233] S. Behnke and N. B. Karayiannis. Cnet: competitive neural trees for pattern classification. In *IEEE International Conference on Neural Networks*, 1996.

[234] M. M. Campos and G. A. Carpenter. S-tree: self-organizing trees for data clustering and online vector quantization. *Neural Networks*, 14:505–525, 2001.

[235] J. Dopazo and J. M. Carazo. Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. *J Mol Evol*, 44:226–33, 1997.

[236] Bernd Fritzke. Growing cell structures - a self-organizing network for unsupervised and supervised learning. *Neural Networks*, 7:1441–1460, 1993.

[237] T.M Martinetz, S.G. Berkovich, and K. Schulten. "neural gas" for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4:558–569, 1993.

[238] Bernd Fritzke. A growing neural gas network learns topologies. In *Advances in Neural Information Processing Systems*, volume 7, pages 625–632, 1995.

[239] Bernd Fritzke. Kohonen feature maps and growing cell structures - a performance comparison. In Stephen Jose Hanson, Jack D. Cowan, and C. Lee Giles, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 123–130, Denver, Colorado, USA, 1993. Morgan Kaufmann.

[240] G. A. Carpenter and S. Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics and Image Processing*, 37:54–115, 1987.

[241] G. A. Carpenter and S. Grossberg. Art2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics*, 26, 1987.

[242] G. A. Carpenter and S. Grossberg. Art3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures. *Neural Networks*, 3:129–152, 1990.

[243] G. A. Carpenter, S. Grossberg, and J. H. Reynolds. Artmap: Supervised real-time learning and classification of non-stationary data by a self-organizing neural network. *Neural Networks*, 4:565–588, 1991.

[244] C. M. Bishop, M. Svensen, and C. K. I. Williams. Gtm: The generative topographic mapping. *Neural Computation*, 10:215–234, 1998.

[245] R. Linsker. How to generate ordered maps by maximizing the mutual information between input and output signals. *Neural Computation*, 1:402–411, 1989.

[246] M. M. Van Hulle. The formation of topographic maps that maximize the average mutual information of the output responses to noiseless input signals. *Neural Computation*, 9:595–607, 1997.

[247] M. M. Van Hulle. *Faithful Representations and Topographic Maps*. Wiley-Interscience, 2000.

[248] Marc M. Van Hulle. Kernel-based equiprobabilistic topographic map formation. *Neural Computation*, 10:1847–1871, 1998.

[249] AM DiGeorge. Congenital absence of the thymus and its immunologic consequences: concurrence with congenital hypoparathyroidism. *White Plains, NY: March of Dimes-Birth Defects Foundation*, IV(1):116–21, 1968.

[250] E. A. Lindsay, R. Goldberg, V. Jurecic, B. Morrow, C. Carlson, R. S. Kucherlapati, R. J. Shprintzen, and A. Baldini. Velo-cardio-facial syndrome: frequency and extent of 22q11 deletions. *Am J Med Genet*, 57(3):514–522, Jul 1995.

[251] J. Burn, A. Takao, D. Wilson, I. Cross, K. Momma, R. Wadey, P. Scambler, and J. Goodship. Conotruncal anomaly face syndrome is associated with a deletion within chromosome 22q11. *J Med Genet*, 30(10):822–824, Oct 1993.

[252] S. Oskarsdttir, M. Vujic, and A. Fasth. Incidence and prevalence of the 22q11 deletion syndrome: a population-based study in western sweden. *Arch Dis Child*, 89(2):148–151, Feb 2004.

[253] L. Edelmann, R. K. Pandita, and B. E. Morrow. Low-copy repeats mediate the common 3-mb deletion in patients with velo-cardio-facial syndrome. *Am J Hum Genet*, 64(4):1076–1086, Apr 1999.

[254] D. M. McDonald-McGinn, R. Kirschner, E. Goldmuntz, K. Sullivan, P. Eicher, M. Gerdes, E. Moss, C. Solot, P. Wang, I. Jacobs, S. Handler, C. Knightly, K. Heher, M. Wilson, J. E. Ming, K. Grace, D. Driscoll, P. Pasquariello, P. Randall, D. Larossa, B. S. Emanuel, and E. H. Zackai. The philadelphia story: the 22q11.2 deletion: report on 250 patients. *Genet Couns*, 10(1):11–24, 1999.

[255] Kathleen E. Sullivan. The clinical, immunological, and molecular spectrum of chromosome 22q11.2 deletion syndrome and digeorge syndrome. *Curr Opin Allergy Clin Immunol*, 4(6):505–512, Dec 2004.

[256] D. M. McDonald-McGinn, D. LaRossa, E. Goldmuntz, K. Sullivan, P. Eicher, M. Gerdes, E. Moss, P. Wang, C. Solot, P. Schultz, D. Lynch, P. Bingham, G. Keenan, S. Weinzimer, J. E. Ming, D. Driscoll, BJ Clark, 3rd, R. Markowitz, A. Cohen, T. Moshang, P. Pasquariello, P. Randall, B. S. Emanuel, and E. H. Zackai. The 22q11.2 deletion: screening, diagnostic workup, and outcome of results; report on 181 patients. *Genet Test*, 1(2):99–108, 1997.

[257] C. B. Solot, C. Knightly, S. D. Handler, M. Gerdes, D. M. McDonald-McGinn, E. Moss, P. Wang, M. Cohen, P. Randall, D. Larossa, and D. A. Driscoll. Communication disorders in the 22q11.2 microdeletion syndrome. *J Commun Disord*, 33(3):187–203; quiz 203–4, 2000.

[258] M. Gerdes, C. Solot, P. P. Wang, E. Moss, D. LaRossa, P. Randall, E. Goldmuntz, BJ Clark, 3rd, D. A. Driscoll, A. Jawad, B. S. Emanuel, D. M. McDonald-McGinn, M. L. Batshaw, and E. H. Zackai. Cognitive and behavior profile of preschool children with chromosome 22q11.2 deletion. *Am J Med Genet*, 85(2):127–133, Jul 1999.

[259] N. J. Scherer, L. L. D'Antonio, and J. H. Kalbfleisch. Early speech and language development in children with velocardiofacial syndrome. *Am J Med Genet*, 88(6):714–723, Dec 1999.

[260] Kate Baker and Jacob A S. Vorstman. Is there a core neuropsychiatric phenotype in 22q11.2 deletion syndrome? *Curr Opin Neurol*, 25(2):131–137, Apr 2012.

[261] Kevin M. Antshel, Wanda Fremont, Nancy J. Roizen, Robert Shprintzen, Anne Marie Higgins, Amit Dhamoon, and Wendy R. Kates. Adhd, major depressive disorder, and simple phobias are prevalent psychiatric conditions in youth with velocardiofacial syndrome. *J Am Acad Child Adolesc Psychiatry*, 45(5):596–603, May 2006.

[262] A. S. Bassett, K. Hodgkinson, E. W. Chow, S. Correia, L. E. Scutt, and R. Weksberg. 22q11 deletion syndrome in adults with schizophrenia. *Am J Med Genet*, 81(4):328–337, Jul 1998.

[263] Anne S. Bassett, Eva W C. Chow, Philip AbdelMalik, Mirona Gheorghiu, Janice Husted, and Rosanna Weksberg. The schizophrenia phenotype in 22q11 deletion syndrome. *Am J Psychiatry*, 160(9):1580–1586, Sep 2003.

[264] Janneke Zinkstok and Thrse van Amelsvoort. Neuropsychological profile and neuroimaging in patients with 22q11.2 deletion syndrome: a review. *Child Neuropsychol*, 11(1):21–37, Feb 2005.

[265] Sarah E. Fine, Alison Weissman, Marsha Gerdes, Jennifer Pinto-Martin, Elaine H. Zackai, Donna M. McDonald-McGinn, and Beverly S. Emanuel. Autism spectrum disorders and symptoms in children with molecularly confirmed 22q11.2 deletion syndrome. *J Autism Dev Disord*, 35(4):461–470, Aug 2005.

[266] Jacob A S. Vorstman, Monique E J. Morcus, Sasja N. Duijff, Petra W J. Klaassen, Josien A. Heineman-de Boer, Frits A. Beemer, Hanna Swaab, Ren S. Kahn, and Herman van Engeland. The 22q11.2 deletion in children: high rate of autistic disorders and early onset of psychotic symptoms. *J Am Acad Child Adolesc Psychiatry*, 45(9):1104–1113, Sep 2006.

[267] H. F. Sutherland, U. J. Kim, and P. J. Scambler. Cloning and comparative mapping of the digeorge syndrome critical region in the mouse. *Genomics*, 52(1):37–43, Aug 1998.

[268] A. Puech, B. Saint-Jore, B. Funke, D. J. Gilbert, H. Sirotkin, N. G. Copeland, N. A. Jenkins, R. Kucherlapati, B. Morrow, and A. I. Skoultchi. Comparative mapping of the human 22q11 chromosomal region and the orthologous region in mice reveals complex changes in gene organization. *Proc Natl Acad Sci U S A*, 94(26):14608–14613, Dec 1997.

[269] E. A. Lindsay, A. Botta, V. Jurecic, S. Carattini-Rivera, Y. C. Cheah, H. M. Rosenblatt, A. Bradley, and A. Baldini. Congenital heart disease in mice deficient for the digeorge syndrome region. *Nature*, 401(6751):379–383, Sep 1999.

[270] A. Puech, B. Saint-Jore, S. Merscher, R. G. Russell, D. Cherif, H. Sirotkin, H. Xu, S. Factor, R. Kucherlapati, and A. I. Skoultchi. Normal cardiovascular development in mice deficient for 16 genes in 550 kb of the velocardiofacial/digeorge syndrome region. *Proc Natl Acad Sci U S A*, 97(18):10090–10095, Aug 2000.

[271] E. A. Lindsay, F. Vitelli, H. Su, M. Morishima, T. Huynh, T. Pramparo, V. Jurecic, G. Ogunrinu, H. F. Sutherland, P. J. Scambler, A. Bradley, and A. Baldini. Tbx1 haploinsufficieny in the digeorge syndrome region causes aortic arch defects in mice. *Nature*, 410(6824):97–101, Mar 2001.

[272] S. Merscher, B. Funke, J. A. Epstein, J. Heyer, A. Puech, M. M. Lu, R. J. Xavier, M. B. Demay, R. G. Russell, S. Factor, K. Tokooya, B. S. Jore, M. Lopez, R. K. Pandita, M. Lia, D. Carrion, H. Xu, H. Schorle, J. B. Kobler, P. Scambler, A. Wynshaw-Boris, A. I. Skoultchi, B. E. Morrow, and R. Kucherlapati. Tbx1 is responsible for cardiovascular defects in velo-cardio-facial/digeorge syndrome. *Cell*, 104(4):619–629, Feb 2001.

[273] L. A. Jerome and V. E. Papaioannou. Digeorge syndrome phenotype in mice mutant for the t-box gene, tbx1. *Nat Genet*, 27(3):286–291, Mar 2001.

[274] Deborah L. Guris, Gregg Duester, Virginia E. Papaioannou, and Akira Imamoto. Dose-dependent interaction of tbx1 and crkl and locally aberrant ra signaling in a model of del22q11 syndrome. *Dev Cell*, 10(1):81–92, Jan 2006.

[275] R. J. Bollag, Z. Siegfried, J. A. Cebra-Thomas, N. Garvey, E. M. Davison, and L. M. Silver. An ancient family of embryonically expressed mouse genes sharing a conserved protein motif with the t locus. *Nat Genet*, 7(3):383–389, Jul 1994.

[276] D. L. Chapman, N. Garvey, S. Hancock, M. Alexiou, S. I. Agulnik, J. J. Gibson-Brown, J. Cebra-Thomas, R. J. Bollag, L. M. Silver, and V. E. Papaioannou. Expression of the t-box family genes, tbx1-tbx5, during early mouse development. *Dev Dyn*, 206(4):379–390, Aug 1996.

[277] Francesca Vitelli, Masae Morishima, Ilaria Taddei, Elizabeth A. Lindsay, and Antonio Baldini. Tbx1 mutation causes multiple cardiovascular defects and disrupts neural crest and cranial nerve migratory pathways. *Hum Mol Genet*, 11(8):915–922, Apr 2002.

[278] Francesca Vitelli, Ilaria Taddei, Masae Morishima, Erik N. Meyers, Elizabeth A. Lindsay, and Antonio Baldini. A genetic link between tbx1 and fibroblast growth factor signaling. *Development*, 129(19):4605–4611, Oct 2002.

[279] Amlie Calmont, Sarah Ivins, Kelly Lammerts Van Bueren, Irinna Papangeli, Vanessa Kyriakopoulou, William D. Andrews, James F. Martin, Anne M. Moon, Elizabeth A. Illingworth, M Albert Basson, and Peter J. Scambler. Tbx1 controls cardiac neural crest cell migration during arch artery development by regulating gbx2 expression in the pharyngeal ectoderm. *Development*, 136(18):3173–3183, Sep 2009.

[280] Ann Marie Scholl and Margaret L. Kirby. Signals controlling neural crest contributions to the heart. *Wiley Interdiscip Rev Syst Biol Med*, 1(2):220–227, 2009.

[281] Margaret L. Kirby and Mary R. Hutson. Factors controlling cardiac neural crest cell migration. *Cell Adh Migr*, 4(4):609–621, 2010.

[282] Lucile Ryckebsch, Nicolas Bertrand, Karim Mesbah, Fanny Bajolle, Karen Niederreither, Robert G. Kelly, and Stphane Zaffran. Decreased levels of embryonic retinoic acid synthesis accelerate recovery from arterial growth delay in a mouse model of digeorge syndrome. *Circ Res*, 106(4):686–694, Mar 2010.

[283] Gregg Duester. Retinoic acid synthesis and signaling during early organogenesis. *Cell*, 134(6):921–931, Sep 2008.

[284] E. D. Dickman, C. Thaller, and S. M. Smith. Temporally-regulated retinoic acid depletion produces specific neural crest, ocular and nervous system defects. *Development*, 124(16):3111–3121, Aug 1997.

[285] E. J. Lammer, D. T. Chen, R. M. Hoar, N. D. Agnish, P. J. Benke, J. T. Braun, C. J. Curry, P. M. Fernhoff, AW Grix, Jr, and I. T. Lott. Retinoic acid embryopathy. *N Engl J Med*, 313(14):837–841, Oct 1985.

[286] V. Garg, C. Yamagishi, T. Hu, I. S. Kathiriya, H. Yamagishi, and D. Srivastava. Tbx1, a digeorge syndrome candidate gene, is regulated by sonic hedgehog during pharyngeal arch development. *Dev Biol*, 235(1):62–73, Jul 2001.

[287] Hiroyuki Yamagishi, Jun Maeda, Tonghuan Hu, John McAnally, Simon J. Conway, Tsutomu Kume, Erik N. Meyers, Chihiro Yamagishi, and Deepak Srivastava. Tbx1 is regulated by tissue-specific forkhead proteins through a common sonic hedgehog-responsive enhancer. *Genes Dev*, 17(2):269–281, Jan 2003.

[288] Richard Paylor, Beate Glaser, Annalisa Mupo, Paris Ataliotis, Corinne Spencer, Angela Sobotka, Chelsey Sparks, Chul-Hee Choi, John Oghalai, Sarah Curran, Kieran C. Murphy, Stephen Monks, Nigel Williams, Michael C. O'Donovan, Michael J. Owen, Peter J. Scambler, and Elizabeth Lindsay. Tbx1 haploinsufficiency is linked to behavioral disorders in mice and humans: implications for 22q11 deletion syndrome. *Proc Natl Acad Sci U S A*, 103(20):7729–7734, May 2006.

[289] Takeshi Hiramoto, Gina Kang, Go Suzuki, Yasushi Satoh, Raju Kucherlapati, Yasuhiro Watanabe, and Noboru Hiroi. Tbx1: identification of a 22q11.2 gene as a risk factor for autism spectrum disorder in a mouse model. *Hum Mol Genet*, 20(24):4775–4785, Dec 2011.

[290] Giovanna Lalli. Extracellular signals controlling neuroblast migration in the postnatal brain. *Adv Exp Med Biol*, 800:149–180, 2014.

[291] Oscar Arias-Carrin. Basic mechanisms of rtms: Implications in parkinson's disease. *Int Arch Med*, 1(1):2, 2008.

[292] Maurice A. Curtis, Monica Kam, Ulf Nannmark, Michelle F. Anderson, Mathilda Zetterstrom Axell, Carsten Wikkelso, Stig Holts, Willeke M C. van Roon-Mom, Thomas Bjrk-Eriksson, Claes Nordborg, Jonas Frisn, Michael Dragunow, Richard L M. Faull, and Peter S. Eriksson. Human neuroblasts migrate to the olfactory bulb via a lateral ventricular extension. *Science*, 315(5816):1243–1249, Mar 2007.

[293] Nader Sanai, Mitchel S. Berger, Jose Manuel Garcia-Verdugo, and Arturo Alvarez-Buylla. Comment on "human neuroblasts migrate to the olfactory bulb via a lateral ventricular extension". *Science*, 318(5849):393; author reply 393, Oct 2007.

[294] C. T. Ekdahl, Z. Kokaia, and O. Lindvall. Brain inflammation and adult neurogenesis: the dual role of microglia. *Neuroscience*, 158(3):1021–1029, Feb 2009.

[295] Toshiaki Nakashiba, Jesse D. Cushman, Kenneth A. Pelkey, Sophie Renaudineau, Derek L. Buhl, Thomas J. McHugh, Vanessa Rodriguez Barrera, Ramesh Chittajallu, Keisuke S. Iwamoto, Chris J. McBain, Michael S. Fanselow, and Susumu Tonegawa. Young dentate granule cells mediate pattern separation, whereas old granule cells facilitate pattern completion. *Cell*, 149(1):188–201, Mar 2012.

[296] James B. Aimone, Janet Wiles, and Fred H. Gage. Potential role for adult neurogenesis in the encoding of time in new memories. *Nat Neurosci*, 9(6):723–727, Jun 2006.

[297] Kathryn M. Harper, Takeshi Hiramoto, Kenji Tanigaki, Gina Kang, Go Suzuki, William Trimble, and Noboru Hiroi. Alterations of social interaction through genetic and environmental manipulation of the 22q11.2 gene sept5 in the mouse brain. *Hum Mol Genet*, 21(15):3489–3499, Aug 2012.

[298] Wendy Adams and Maarten van den Buuse. Hippocampal serotonin depletion facilitates the enhancement of prepulse inhibition by risperidone: possible role of 5-ht(2c) receptors in the dorsal hippocampus. *Neuropharmacology*, 61(3):458–467, Sep 2011.

[299] R. Muramatsu, Y. Ikegaya, N. Matsuki, and R. Koyama. Neonatally born granule cells numerically dominate adult mice dentate gyrus. *Neuroscience*, 148(3):593–598, Sep 2007.

[300] G. Kempermann and F. H. Gage. Neurogenesis in the adult hippocampus. *Novartis Found Symp*, 231:220–35; discussion 235–41, 302–6, 2000.

[301] Cory Y. McLean, Dave Bristor, Michael Hiller, Shoa L. Clarke, Bruce T. Schaar, Craig B. Lowe, Aaron M. Wenger, and Gill Bejerano. Great improves functional interpretation of cis-regulatory regions. *Nat Biotechnol*, 28(5):495–501, May 2010.

[302] Li-Ming Xu, Jia-Rui Li, Yue Huang, Min Zhao, Xing Tang, and Liping Wei. Autismkb: an evidence-based knowledgebase of autism genetics. *Nucleic Acids Res*, 40(Database issue):D1016–D1022, Jan 2012.

[303] Saumyendra N. Basu, Ravi Kollu, and Sharmila Banerjee-Basu. Autdb: a gene reference resource for autism research. *Nucleic Acids Res*, 37(Database issue):D832–D836, Jan 2009.

[304] Nicole C. Allen, Sachin Bagade, Matthew B. McQueen, John P A. Ioannidis, Fotini K. Kavvoura, Muin J. Khoury, Rudolph E. Tanzi, and Lars Bertram. Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the szgene database. *Nat Genet*, 40(7):827–834, Jul 2008.

[305] P. Jia, J. Sun, A. Y. Guo, and Z. Zhao. Szgr: a comprehensive schizophrenia gene resource. *Mol Psychiatry*, 15(5):453–462, May 2010.

[306] Jingchun Sun, Po-Hsiu Kuo, Brien P. Riley, Kenneth S. Kendler, and Zhongming Zhao. Candidate genes for schizophrenia: a survey of association studies and gene ranking. *Am J Med Genet B Neuropsychiatr Genet*, 147B(7):1173–1181, Oct 2008.

[307] Rupert W. Overall, Maciej Paszkowski-Rogacz, and Gerd Kempermann. The mammalian adult neurogenesis gene ontology (mango) provides a structural framework for published information on genes regulating adult hippocampal neurogenesis. *PLoS One*, 7(11):e48527, 2012.

[308] Nir Oksenberg and Nadav Ahituv. The role of auts2 in neurodevelopment and human evolution. *Trends Genet*, 29(10):600–608, Oct 2013.

[309] Razia Sultana, Chang-En Yu, Jun Yu, Jeffery Munson, Donghui Chen, Wenhui Hua, Annette Estes, Fanny Cortes, Flora de la Barra, Dongmei Yu, Syed T. Haider, Barbara J. Trask, Eric D. Green, Wendy H. Raskind, Christine M. Disteche, Ellen Wijsman, Geraldine Dawson, Daniel R. Storm, Gerard D. Schellenberg, and Enrique C. Villacres. Identification of a novel gene on chromosome 7q11.2 interrupted by a translocation breakpoint in a pair of autistic twins. *Genomics*, 80(2):129–134, Aug 2002.

[310] M. L. Hamshere, E. K. Green, I. R. Jones, L. Jones, V. Moskvina, G. Kirov, D. Grozeva, I. Nikolov, D. Vukcevic, S. Caesar, K. Gordon-Smith, C. Fraser, E. Russell, G. Breen, D. St Clair, D. A. Collier, A. H. Young, I. N. Ferrier, A. Farmer, P. McGuffin, Wellcome Trust Case Control Consortium , P. A. Holmans, M. J. Owen, M. C. O'Donovan, and N. Craddock. Genetic utility of broadly defined bipolar schizoaffective disorder as a diagnostic concept. *Br J Psychiatry*, 195(1):23–29, Jul 2009.

[311] Eiji Hattori, Tomoko Toyota, Yuichi Ishitsuka, Yoshimi Iwayama, Kazuo Yamada, Hiroshi Ujike, Yukitaka Morita, Masafumi Kodama, Kenji Nakata, Yoshio Minabe, Kazuhiko Nakamura, Yasuhide Iwata, Nori Takei, Norio Mori, Hiroshi Naitoh, Yoshio Yamanouchi, Nakao Iwata, Norio Ozaki, Tadafumi Kato, Toru Nishikawa, Atsushi Kashiwa, Mika Suzuki, Kunihiko Shioe, Manabu Shinohara, Masami Hirano, Shinichiro Nanko, Akihisa Akahane, Mikako Ueno, Naoshi Kaneko, Yuichiro Watanabe, Toshiyuki Someya, Kenji Hashimoto, Masaomi Iyo, Masanari Itokawa, Makoto Arai, Masahiro Nankai, Toshiya Inada, Sumiko Yoshida, Hiroshi Kunugi, Michiko Nakamura, Yoshimi Iijima, Yuji Okazaki, Teruhiko Higuchi, and Takeo Yoshikawa. Preliminary genome-wide association study of bipolar disorder in the japanese population. *Am J Med Genet B Neuropsychiatr Genet*, 150B(8):1110–1117, Dec 2009.

[312] Michael E. Talkowski, Jill A. Rosenfeld, Ian Blumenthal, Vamsee Pillalamarri, Colby Chiang, Adrian Heilbut, Carl Ernst, Carrie Hanscom, Elizabeth Rossin, Amelia M. Lindgren, Shahrin Pereira, Douglas Ruderfer, Andrew Kirby, Stephan Ripke, David J. Harris, Ji-Hyun Lee, Kyungsoo Ha, Hyung-Goo Kim, Benjamin D. Solomon, Andrea L. Gropman, Diane Lucente, Katherine Sims, Toshiro K. Ohsumi, Mark L. Borowsky, Stephanie Loranger, Bradley Quade, Kasper Lage, Judith Miles, Bai-Lin Wu, Yiping Shen, Benjamin Neale, Lisa G. Shaffer, Mark J. Daly, Cynthia C. Morton, and James F. Gusella. Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell*, 149(3):525–537, Apr 2012.

[313] Santhosh Girirajan, Zoran Brkanac, Bradley P. Coe, Carl Baker, Laura Vives, Tiffany H. Vu, Neil Shafer, Raphael Bernier, Giovanni B. Ferrero, Margherita Silengo, Stephen T. Warren, Carlos S. Moreno, Marco Fichera, Corrado Romano, Wendy H. Raskind, and Evan E. Eichler. Relative burden of large cnvs on a range of neurodevelopmental phenotypes. *PLoS Genet*, 7(11):e1002334, Nov 2011.

[314] J. Elia, X. Gai, H. M. Xie, J. C. Perin, E. Geiger, J. T. Glessner, M. D'arcy, R. deBerardinis, E. Frackelton, C. Kim, F. Lantieri, B. M. Muganga, L. Wang, T. Takeda, E. F. Rappaport, S F A. Grant, W. Berrettini, M. Devoto, T. H. Shaikh, H. Hakonarson, and P. S. White. Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. *Mol Psychiatry*, 15(6):637–646, Jun 2010.

[315] Heather C. Mefford, Hiltrud Muhle, Philipp Ostertag, Sarah von Spiczak, Karen Buysse, Carl Baker, Andre Franke, Alain Malafosse, Pierre Genton, Pierre Thomas, Christina A. Gurnett, Stefan Schreiber, Alexander G. Bassuk, Michel Guipponi, Ulrich Stephani, Ingo Helbig, and Evan E. Eichler. Genome-wide copy number variation in epilepsy: novel susceptibility loci in idiopathic generalized and focal epilepsies. *PLoS Genet*, 6(5):e1000962, May 2010.

[316] A-M. Lepagnol-Bestel, G. Maussion, B. Boda, A. Cardona, Y. Iwayama, A-L. Delezoide, J-M. Moalic, D. Muller, B. Dean, T. Yoshikawa, P. Gorwood, J. D. Buxbaum, N. Ramoz, and M. Simonneau. Slc25a12 expression is associated with neurite outgrowth and is upregulated in the prefrontal cortex of autistic subjects. *Mol Psychiatry*, 13(4):385–397, Apr 2008.

[317] Francesco Bedogni, Rebecca D. Hodge, Branden R. Nelson, Erika A. Frederick, Naoko Shiba, Ray A. Daza, and Robert F. Hevner. Autism susceptibility candidate 2 (auts2) encodes a nuclear protein expressed in developing brain regions implicated in autism neuropathology. *Gene Expr Patterns*, 10(1):9–15, Jan 2010.

[318] D. St Clair, D. Blackwood, W. Muir, A. Carothers, M. Walker, G. Spowart, C. Gosden, and H. J. Evans. Association within a family of a balanced autosomal translocation with major mental illness. *Lancet*, 336(8706):13–16, Jul 1990.

[319] J Kirsty Millar, Rachel James, Nicholas J. Brandon, and Pippa A. Thomson. Disc1 and disc2: discovering and dissecting molecular mechanisms underlying psychiatric illness. *Ann Med*, 36(5):367–378, 2004.

[320] Yann Le Strat, Nicolas Ramoz, and Philip Gorwood. The role of genes involved in neuroplasticity and neurogenesis in the observation of a gene-environment interaction (gxe) in schizophrenia. *Curr Mol Med*, 9(4):506–518, May 2009.

[321] Ju Young Kim, Xin Duan, Cindy Y. Liu, Mi-Hyeon Jang, Junjie U. Guo, Nattapol Pow-anpongkul, Eunchai Kang, Hongjun Song, and Guo-li Ming. Disc1 regulates new neuron development in the adult brain via modulation of akt-mtor signaling through kiaa1212. *Neuron*, 63(6):761–773, Sep 2009.

[322] Yingwei Mao, Xuecai Ge, Christopher L. Frank, Jon M. Madison, Angela N. Koehler, Mary Kathryn Doud, Carlos Tassa, Erin M. Berry, Takahiro Soda, Karun K. Singh, Travis Biechele, Tracey L. Petryshen, Randall T. Moon, Stephen J. Haggarty, and Li-Huei Tsai. Disrupted in schizophrenia 1 regulates neuronal progenitor proliferation via modulation of gsk3beta/beta-catenin signaling. *Cell*, 136(6):1017–1031, Mar 2009.

[323] Xin Duan, Jay H. Chang, Shaoyu Ge, Regina L. Faulkner, Ju Young Kim, Yasuji Kitabatake, Xiao-bo Liu, Chih-Hao Yang, J Dedrick Jordan, Dengke K. Ma, Cindy Y. Liu, Sundar Ganesan, Hwai-Jong Cheng, Guo-li Ming, Bai Lu, and Hongjun Song. Disrupted-in-schizophrenia 1 regulates integration of newly generated neurons in the adult brain. *Cell*, 130(6):1146–1158, Sep 2007.

[324] Fanfan Zheng, Lifang Wang, Meixiang Jia, Weihua Yue, Yan Ruan, Tianlan Lu, Jing Liu, Jun Li, and Dai Zhang. Evidence for association between disrupted-in-schizophrenia 1 (disc1) gene polymorphisms and autism in chinese han population: a family-based association study. *Behav Brain Funct*, 7:14, 2011.

[325] H. Kilpinen, T. Ylisaukko-Oja, W. Hennah, O. M. Palo, T. Varilo, R. Vanhala, T. Nieminen-von Wendt, L. von Wendt, T. Paunio, and L. Peltonen. Association of disc1 with autism and asperger syndrome. *Mol Psychiatry*, 13(2):187–196, Feb 2008.

[326] A. L. Joyner. Engrailed, wnt and pax genes regulate midbrain–hindbrain development. *Trends Genet*, 12(1):15–20, Jan 1996.

[327] K. J. Millen, W. Wurst, K. Herrup, and A. L. Joyner. Abnormal embryonic cerebellar development and patterning of postnatal foliation in two mouse engrailed-2 mutants. *Development*, 120(3):695–706, Mar 1994.

[328] C. A. Davis, S. E. Noble-Topham, J. Rossant, and A. L. Joyner. Expression of the homeo box-containing gene en-2 delineates a specific region of the developing mouse brain. *Genes Dev*, 2(3):361–371, Mar 1988.

[329] P. P. Tripathi, P. Sgad, M. Scali, C. Viaggi, S. Casarosa, H. H. Simon, F. Vaglini, G. U. Corsini, and Y. Bozzi. Increased susceptibility to kainic acid-induced seizures in engrailed-2 knockout mice. *Neuroscience*, 159(2):842–849, Mar 2009.

[330] Ekrem Maloku, Ignacio R. Covelo, Ingeborg Hanbauer, Alessandro Guidotti, Bashkim Kadriu, Qiaoyan Hu, John M. Davis, and Erminio Costa. Lower number of cerebellar purkinje neurons in psychosis is associated with reduced reelin expression. *Proc Natl Acad Sci U S A*, 107(9):4407–4411, Mar 2010.

[331] Michelle A. Cheh, James H. Millonig, Lauren M. Roselli, Xue Ming, Erin Jacobsen, Silky Kamdar, and George C. Wagner. En2 knockout mice display neurobehavioral and neurochemical alterations relevant to autism spectrum disorder. *Brain Res*, 1116(1):166–176, Oct 2006.

[332] M. Genestine, L. Lin, Y. Yan, S. Prem, J. H. Millonig, and E. DiCicco-Bloom. Absence of engrailed 2 (en2), the autism spectrum disorder (asd) associated gene, alters locus coeruleus fiber elaboration and hippocampal neurogenesis and apoptosis. In *International Meeting for Autism Research*, 2011.

[333] N. Gharani, R. Benayed, V. Mancuso, L. M. Brzustowicz, and J. H. Millonig. Association of the homeobox transcription factor, engrailed 2, 3, with autism spectrum disorder. *Mol Psychiatry*, 9(5):474–484, May 2004.

[334] E. Petit, J. Hrault, J. Martineau, A. Perrot, C. Barthlmy, L. Hameury, D. Sauvage, G. Lelord, and J. P. Mh. Association study with two markers of a human homeogene in infantile autism. *J Med Genet*, 32(4):269–274, Apr 1995.

[335] B. Sen, A Surindro Singh, S. Sinha, A. Chatterjee, S. Ahmed, S. Ghosh, and R. Usha. Family-based studies indicate association of engrailed 2 gene with autism in an indian population. *Genes Brain Behav*, 9(2):248–255, Mar 2010.

[336] Ida Rissling, Konstantin Strauch, Christine Hft, Wolfgang Hermann Oertel, and Jens Carsten Mller. Haplotype analysis of the engrailed-2 gene in young-onset parkinson's disease. *Neurodegener Dis*, 6(3):102–105, 2009.

[337] Margaret S. Ho, Pei-I. Tsai, and Cheng-Ting Chien. F-box proteins: the key to protein degradation. *J Biomed Sci*, 13(2):181–191, Mar 2006.

[338] Joseph T. Glessner, Kai Wang, Guiqing Cai, Olena Korvatska, Cecilia E. Kim, Shawn Wood, Haitao Zhang, Annette Estes, Camille W. Brune, Jonathan P. Bradfield, Marcin Imielinski, Edward C. Frackelton, Jennifer Reichert, Emily L. Crawford, Jeffrey Munson, Patrick M A. Sleiman, Rosetta Chiavacci, Kiran Annaiah, Kelly Thomas, Cuiping Hou, Wendy Glaberson, James Flory, Frederick Otieno, Maria Garris, Latha Soorya, Lambertus Klei, Joseph Piven, Kacie J. Meyer, Evdokia Anagnostou, Takeshi Sakurai, Rachel M. Game, Danielle S. Rudd, Danielle Zurawiecki, Christopher J. McDougle, Lea K. Davis, Judith Miller, David J. Posey, Shana Michaels, Alexander Kolevzon, Jeremy M. Silverman, Raphael Bernier, Susan E. Levy, Robert T. Schultz, Geraldine Dawson, Thomas Owley, William M. McMahon, Thomas H. Wassink, John A. Sweeney, John I. Nurnberger, Hilary Coon, James S. Sutcliffe, Nancy J. Minshew, Struan F A. Grant, Maja Bucan, Edwin H. Cook, Joseph D. Buxbaum, Bernie Devlin, Gerard D. Schellenberg, and Hakon Hakonarson. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature*, 459(7246):569–573, May 2009.

[339] D R H. de Bruijn, A H A. van Dijk, R. Pfundt, A. Hoischen, G F M. Merkx, G. A. Gradek, H. Lybk, A. Stray-Pedersen, H. G. Brunner, and G. Houge. Severe progressive autism associated with two de novo changes: A 2.6-mb 2q31.1 deletion and a balanced t(14;21)(q21.1;p11.2) translocation with long-range epigenetic silencing of lrfn5 expression. *Mol Syndromol*, 1(1):46–57, Feb 2010.

[340] Jillian P. Casey, Tiago Magalhaes, Judith M. Conroy, Regina Regan, Naisha Shah, Richard Anney, Denis C. Shields, Brett S. Abrahams, Joana Almeida, Elena Bacchelli, Anthony J. Bailey, Gillian Baird, Agatino Battaglia, Tom Berney, Nadia Bolshakova, Patrick F. Bolton, Thomas Bourgeron, Sean Brennan, Phil Cali, Catarina Correia, Christina Corsello, Marc Coutanche, Geraldine Dawson, Maretha de Jonge, Richard Delorme, Eftichia Duketis, Frederico Duque, Annette Estes, Penny Farrar, Bridget A. Fernandez, Susan E. Folstein, Suzanne Foley, Eric Fombonne, Christine M. Freitag, John Gilbert, Christopher Gillberg, Joseph T. Glessner, Jonathan Green, Stephen J. Guter, Hakon Hakonarson, Richard Holt, Gillian Hughes, Vanessa Hus, Roberta Igliozzi, Cecilia Kim, Sabine M. Klauck, Alexander Kolevzon, Janine A. Lamb, Marion Leboyer, Ann Le Couteur, Bennett L. Leventhal, Catherine Lord, Sabata C. Lund, Elena Maestrini, Carine Mantoulan, Christian R. Marshall, Helen McConachie, Christopher J. McDougle, Jane McGrath, William M. McMahon, Alison Merikangas, Judith Miller, Fiorella Minopoli, Ghazala K. Mirza, Jeff Munson, Stanley F. Nelson, Gudrun Nygren, Guiomar Oliveira, Alistair T. Pagnamenta, Katerina Papanikolaou, Jeremy R. Parr, Barbara Parrini, Andrew Pickles, Dalila Pinto, Joseph Piven, David J. Posey, Annemarie Poustka, Fritz Poustka, Jiannis Ragoussis, Bernadette Roge, Michael L. Rutter, Ana F. Sequeira, Latha Soorya, Ins Sousa, Nuala Sykes, Vera Stoppioni, Raffaella Tancredi, Mat Tauber, Ann P. Thompson, Susanne Thomson, John Tsiantis, Herman Van Engeland, John B. Vincent, Fred Volkmar, Jacob A S. Vorstman, Simon Wallace, Kai Wang, Thomas H. Wassink, Kathy White, Kirsty Wing, Kerstin Wittemeyer, Brian L. Yaspan, Lonnie Zwaigenbaum, Catalina Betancur, Joseph D. Buxbaum, Rita M. Cantor, Edwin H. Cook, Hilary Coon, Michael L. Cuccaro, Daniel H. Geschwind, Jonathan L. Haines, Joachim Hallmayer, Anthony P. Monaco, John I Nurnberger, Jr, Margaret A. Pericak-Vance, Gerard D. Schellenberg, Stephen W. Scherer, James S. Sutcliffe, Peter Szatmari, Veronica J. Vieland, Ellen M. Wijsman, Andrew Green, Michael Gill, Louise Gallagher, Astrid Vicente, and Sean Ennis. A novel approach of homozygous haplotype sharing identifies candidate genes in autism spectrum disorder. *Hum Genet*, 131(4):565–579, Apr 2012.

[341] Erin B. Kaminsky, Vineith Kaul, Justin Paschall, Deanna M. Church, Brian Bunke, Dawn Kunig, Daniel Moreno-De-Luca, Andres Moreno-De-Luca, Jennifer G. Mulle, Stephen T. Warren, Gabriele Richard, John G. Compton, Amy E. Fuller, Troy J. Gliem, Shuwen Huang, Morag N. Collinson, Sarah J. Beal, Todd Ackley, Diane L. Pickering, Denae M. Golden, Emily Aston, Heidi Whitby, Shashirekha Shetty, Michael R. Rossi, M Katharine Rudd, Sarah T. South, Arthur R. Brothman, Warren G. Sanger, Ramaswamy K. Iyer, John A. Crolla, Erik C. Thorland, Swaroop Aradhya, David H. Ledbetter, and Christa L. Martin. An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. *Genet Med*, 13(9):777–784, Sep 2011.

[342] Warren D. Flood, Robert W. Moyer, Anna Tsykin, Grant R. Sutherland, and Simon A. Koblar. Nxf and fbxo33: novel seizure-responsive genes in mice. *Eur J Neurosci*, 20(7):1819–1826, Oct 2004.

[343] T. Maehama and J. E. Dixon. The tumor suppressor, pten/mmac1, dephosphorylates the lipid second messenger, phosphatidylinositol 3,4,5-trisphosphate. *J Biol Chem*, 273(22):13375–13378, May 1998.

[344] Zhenbang Chen, Lloyd C. Trotman, David Shaffer, Hui-Kuan Lin, Zohar A. Dotan, Masaru Niki, Jason A. Koutcher, Howard I. Scher, Thomas Ludwig, William Gerald, Carlos Cordon-Cardo, and Pier Paolo Pandolfi. Crucial role of p53-dependent cellular senescence in suppression of pten-deficient tumorigenesis. *Nature*, 436(7051):725–730, Aug 2005.

[345] Laura Maria Pradella, Cecilia Evangelisti, Claudia Ligorio, Claudio Ceccarelli, Iria Neri, Roberta Zuntini, Laura Benedetta Amato, Simona Ferrari, Alberto Maria Martelli, Giuseppe Gasparre, and Daniela Turchetti. A novel deleterious pten mutation in a patient with early-onset bilateral breast cancer. *BMC Cancer*, 14:70, 2014.

[346] S. I. Wang, J. Puc, J. Li, J. N. Bruce, P. Cairns, D. Sidransky, and R. Parsons. Somatic mutations of pten in glioblastoma multiforme. *Cancer Res*, 57(19):4183–4186, Oct 1997.

[347] KM Lloyd and M. Dennis. Cowden's disease. a possible new symptom complex with multiple system involvement. *Ann Intern Med*, 58:136–142, Jan 1963.

[348] Philippe Buisson, Marc-David Leclair, Sbastien Jacquemont, Guillaume Podevin, Caroline Camby, Albert David, and Yves Heloury. Cutaneous lipoma in children: 5 cases with bannayan-riley-ruvalcaba syndrome. *J Pediatr Surg*, 41(9):1601–1603, Sep 2006.

[349] M. G. Butler, M. J. Dasouki, X-P. Zhou, Z. Talebizadeh, M. Brown, T. N. Takahashi, J. H. Miles, C. H. Wang, R. Stratton, R. Pilarski, and C. Eng. Subset of individuals with autism spectrum disorders and extreme macrocephaly associated with germline pten tumour suppressor gene mutations. *J Med Genet*, 42(4):318–321, Apr 2005.

[350] Kim L. McBride, Elizabeth A. Varga, Matthew T. Pastore, Thomas W. Prior, Kandamurugu Manickam, Joan F. Atkin, and Gail E. Herman. Confirmation study of pten mutations among individuals with autism or developmental delays/mental retardation and macrocephaly. *Autism Res*, 3(3):137–141, Jun 2010.

[351] Elizabeth A. Varga, Matthew Pastore, Thomas Prior, Gail E. Herman, and Kim L. McBride. The prevalence of pten mutations in a clinical pediatric cohort with autism spectrum disorders, developmental delay, and macrocephaly. *Genet Med*, 11(2):111–117, Feb 2009.

[352] Chang-Hyuk Kwon, Bryan W. Luikart, Craig M. Powell, Jing Zhou, Sharon A. Matheny, Wei Zhang, Yanjiao Li, Suzanne J. Baker, and Luis F. Parada. Pten regulates neuronal arborization and social interaction in mice. *Neuron*, 50(3):377–388, May 2006.

[353] M. Groszer, R. Erickson, D. D. Scripture-Adams, R. Lesche, A. Trumpp, J. A. Zack, H. I. Kornblum, X. Liu, and H. Wu. Negative regulation of neural stem/progenitor cell proliferation by the pten tumor suppressor gene in vivo. *Science*, 294(5549):2186–2189, Dec 2001.

[354] Caroline Gregorian, Jonathan Nakashima, Janel Le Belle, John Ohab, Rachel Kim, Annie Liu, Kate Barzan Smith, Matthias Groszer, A Denise Garcia, Michael V. Sofroniew, S Thomas Carmichael, Harley I. Kornblum, Xin Liu, and Hong Wu. Pten deletion in adult neural stem/progenitor cells enhances constitutive neurogenesis. *J Neurosci*, 29(6):1874–1886, Feb 2009.

[355] Anahita Amiri, Woosung Cho, Jing Zhou, Shari G. Birnbaum, Christopher M. Sinton, Rene M. McKay, and Luis F. Parada. Pten deletion in adult hippocampal neural stem/progenitor cells causes cellular abnormalities and alters neurogenesis. *J Neurosci*, 32(17):5880–5890, Apr 2012.

[356] Berten P G M. Ceulemans, Lieve R F. Claes, and Lieven G. Lagae. Clinical correlations of mutations in the scn1a gene: from febrile seizures to severe myoclonic epilepsy in infancy. *Pediatr Neurol*, 30(4):236–243, Apr 2004.

[357] Tateki Fujiwara, Takashi Sugawara, Emi Mazaki-Miyazaki, Yukitoshi Takahashi, Katsuyuki Fukushima, Masako Watanabe, Keita Hara, Tateki Morikawa, Kazuichi Yagi, Kazuhiro Yamakawa, and Yushi Inoue. Mutations of sodium channel alpha subunit type 1 (scn1a) in intractable childhood epilepsies with frequent generalized tonic-clonic seizures. *Brain*, 126(Pt 3):531–546, Mar 2003.

[358] Charlotte Dravet. Dravet syndrome history. *Dev Med Child Neurol*, 53 Suppl 2:1–6, Apr 2011.

[359] Daria Riva, Chiara Vago, Chiara Pantaleoni, Sara Bulgheroni, Massimo Mantegazza, and Silvana Franceschetti. Progressive neurocognitive decline in two children with dravet syndrome, de novo scn1a truncations and different epileptic phenotypes. *Am J Med Genet A*, 149A(10):2339–2345, Oct 2009.

[360] Rima Nabbout and Olivier Dulac. Epileptic encephalopathies: a brief overview. *J Clin Neurophysiol*, 20(6):393–397, 2003.

[361] L. A. Weiss, A. Escayg, J. A. Kearney, M. Trudeau, B. T. MacDonald, M. Mori, J. Reichert, J. D. Buxbaum, and M. H. Meisler. Sodium channels scn1a, scn2a and scn3a in familial autism. *Mol Psychiatry*, 8(2):186–194, Feb 2003.

[362] Stephan J. Sanders, Michael T. Murtha, Abha R. Gupta, John D. Murdoch, Melanie J. Raubeson, A Jeremy Willsey, A Gulhan Ercan-Sencicek, Nicholas M. DiLullo, Neelroop N. Parikshak, Jason L. Stein, Michael F. Walker, Gordon T. Ober, Nicole A. Teran, Youeun Song, Paul El-Fishawy, Ryan C. Murtha, Murim Choi, John D. Overton, Robert D. Bjornson, Nicholas J. Carriero, Kyle A. Meyer, Kaya Bilguvar, Shrikant M. Mane, Nenad Sestan, Richard P. Lifton, Murat Gnel, Kathryn Roeder, Daniel H. Geschwind, Bernie Devlin, and Matthew W. State. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, 485(7397):237–241, May 2012.

[363] A. Bulfone, S. M. Smiga, K. Shimamura, A. Peterson, L. Puelles, and J. L. Rubenstein. T-brain-1: a homolog of brachyury whose expression defines molecularly distinct domains within the cerebral cortex. *Neuron*, 15(1):63–78, Jul 1995.

[364] Chris Englund, Andy Fink, Charmaine Lau, Diane Pham, Ray A M. Daza, Alessandro Bulfone, Tom Kowalczyk, and Robert F. Hevner. Pax6, tbr2, and tbr1 are expressed sequentially by radial glia, intermediate progenitor cells, and postmitotic neurons in developing neocortex. *J Neurosci*, 25(1):247–251, Jan 2005.

[365] R. F. Hevner, L. Shi, N. Justice, Y. Hsueh, M. Sheng, S. Smiga, A. Bulfone, A. M. Goffinet, A. T. Campagnoni, and J. L. Rubenstein. Tbr1 regulates differentiation of the preplate and layer 6. *Neuron*, 29(2):353–366, Feb 2001.

[366] S Hossein Fatemi, Anne V. Snow, Joel M. Stary, Mohsen Araghi-Niknam, Teri J. Reutiman, Suzanne Lee, Andrew I. Brooks, and David A. Pearce. Reelin signaling is impaired in autism. *Biol Psychiatry*, 57(7):777–787, Apr 2005.

[367] Laurent Roybon, Tomas Deierborg, Patrik Brundin, and Jia-Yi Li. Involvement of ngn2, tbr and neurod proteins during postnatal olfactory bulb neurogenesis. *Eur J Neurosci*, 29(2):232–243, Jan 2009.

[368] Yi-Ping Yan, Bradley T. Lang, Raghu Vemuganti, and Robert J. Dempsey. Osteopontin is a mediator of the lateral migration of neuroblasts from the subventricular zone after focal cerebral ischemia. *Neurochem Int*, 55(8):826–832, Dec 2009.

[369] Won Sub Kang, Jin Kyung Park, Su Kang Kim, Hae Jeong Park, Sang Min Lee, Ji Young Song, Joo-Ho Chung, and Jong Woo Kim. Genetic variants of gria1 are associated with susceptibility to schizophrenia in korean population. *Mol Biol Rep*, 39(12):10697–10703, Dec 2012.

[370] K. Wakabayashi, M. Narisawa-Saito, Y. Iwakura, T. Arai, K. Ikeda, H. Takahashi, and H. Nawa. Phenotypic down-regulation of glutamate receptor subunit glur1 in alzheimer's disease. *Neurobiol Aging*, 20(3):287–295, 1999.

[371] Berit Kerner, Anna J. Jasinska, Joseph DeYoung, Maricel Almonte, Oi-Wa Choi, and Nelson B. Freimer. Polymorphisms in the gria1 gene region in psychotic bipolar disorder. *Am J Med Genet B Neuropsychiatr Genet*, 150B(1):24–32, Jan 2009.

[372] T. L. Babb, G. W. Mathern, J. P. Leite, J. K. Pretorius, K. M. Yeoman, and P. A. Kuhlman. Glutamate ampa receptors in the fascia dentata of human and kainate rat hippocampal epilepsy. *Epilepsy Res*, 26(1):193–205, Dec 1996.

[373] Nikola A. Bowden, Rodney J. Scott, and Paul A. Tooney. Altered gene expression in the superior temporal gyrus in schizophrenia. *BMC Genomics*, 9:199, 2008.

[374] Valerie W. Hu and Yinglei Lai. Developing a predictive gene classifier for autism spectrum disorders based upon differential gene expression profiles of phenotypic subgroups. *N Am J Med Sci (Boston)*, 6(3), 2013.

[375] Elliott Rees, James T R. Walters, Kimberly D. Chambert, Colm O'Dushlaine, Jin Szatkiewicz, Alexander L. Richards, Lyudmila Georgieva, Gerwyn Mahoney-Davies, Sophie E. Legge, Jennifer L. Moran, Giulio Genovese, Douglas Levinson, Derek W. Morris, Paul Cormican, Kenneth S. Kendler, Francis A. O'Neill, Brien Riley, Michael Gill, Aiden Corvin, Wellcome Trust Case Control Consortium , Pamela Sklar, Christina Hultman, Carlos Pato, Michele Pato, Patrick F. Sullivan, Pablo V. Gejman, Steven A. McCarroll, Michael C. O'Donovan, Michael J. Owen, and George Kirov. Cnv analysis in a large schizophrenia sample implicates deletions at 16p12.1 and slc1a1 and duplications at 1p36.33 and cgnl1. *Hum Mol Genet*, Nov 2013.

[376] Yongchao Liu, Bertil Schmidt, Weiguo Liu, and Douglas L. Maskell. Cuda-meme: Accelerating motif discovery in biological sequences using cuda-enabled graphics processing units. *Pattern Recogn. Lett.*, 31(14):2170–2177, October 2010.

[377] Maria Luisa Scattoni, Shruti U. Gandhy, Laura Ricceri, and Jacqueline N. Crawley. Unusual repertoire of vocalizations in the btbr t+tf/j mouse model of autism. *PLoS One*, 3(8):e3067, 2008.

[378] Claude E Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27 (3):379423, 1948.

[379] Jasmine M S. Grimsley, Jessica J M. Monaghan, and Jeffrey J. Wenstrup. Development of social vocalizations in mice. *PLoS One*, 6(3):e17460, 2011.

[380] Jedol Dayou, Ng Chee Han, Ho Chong Mun, and Abdul Hamid Ahmad. Classification and identification of frogs sounds based on entropy approach. In *IPCBEE*, volume 3, 2011.

[381] Laurance R. Doyle, Brenda McCowan, Simon Johnston, and Sean F. Hanser. Information theory, animal communication, and the search for extraterrestrial intelligence. *Acta Astronautica*, 68(34):406 – 417, 2011. {SETI} Special Edition.

[382] Herman Wold. *Estimation of Principal Components and Related Models by Iterative Least squares*, pages 391–420. Academic Press, New York, 1966.

[383] Herman Wold. *Encyclopedia of statistical sciences 6*, chapter Partial least squares, page 581591. Wiley,New York, 1985.

[384] Joana Carrola, Cludia M. Rocha, Antnio S. Barros, Ana M. Gil, Brian J. Goodfellow, Isabel M. Carreira, Joo Bernardo, Ana Gomes, Vitor Sousa, Lina Carvalho, and Iola F. Duarte. Metabolic signatures of lung cancer in biofluids: Nmr-based metabonomics of urine. *J Proteome Res*, 10(1):221–230, Jan 2011.

[385] Yan-Gan Chen, Yue-Lin Song, Ying Wang, Yun-Fei Yuan, Xiao-Jun Huang, Wen-Cai Ye, Yi-Tao Wang, and Qing-Wen Zhang. Metabolic differentiations of pueraria lobata and pueraria thomsonii using h nmr spectroscopy and multivariate statistical analysis. *J Pharm Biomed Anal*, 93:51–58, May 2014.

[386] Ana Mara Gmez-Caravaca, Vito Verardo, Annachiara Berardinelli, Emanuele Marconi, and Maria Fiorenza Caboni. A chemometric approach to determine the phenolic compounds in different barley samples by two different stationary phases: A comparison between c18 and pentafluorophenyl core shell columns. *J Chromatogr A*, Jun 2014.

[387] Kim-Anh L Cao, Simon Boitard, and Philippe Besse. Sparse pls discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, 12:253, 2011.

[388] Brooke Sinderberry, Scott Brown, Peter Hammond, Angela F. Stevens, Ulrich Schall, Declan G M. Murphy, Kieran C. Murphy, and Linda E. Campbell. Subtypes in 22q11.2 deletion syndrome associated with behaviour and neurofacial morphology. *Res Dev Disabil*, 34(1):116–125, Jan 2013.

[389] Kathleen Angkustsiri, Beth Goodlin-Jones, Lesley Deprey, Khyati Brahmbhatt, Susan Harris, and Tony J. Simon. Social impairments in chromosome 22q11.2 deletion syndrome (22q11.2ds): autism spectrum disorder or a different endophenotype? *J Autism Dev Disord*, 44(4):739–746, Apr 2014.

[390] C. Lord, M. Rutter, S. Goode, J. Heemsbergen, H. Jordan, L. Mawhood, and E. Schopler. Autism diagnostic observation schedule: a standardized observation of communicative and social behavior. *J Autism Dev Disord*, 19(2):185–212, Jun 1989.

[391] C. Lord, S. Risi, L. Lambrecht, EH Cook, Jr, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter. The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J Autism Dev Disord*, 30(3):205–223, Jun 2000.

[392] Rhiannon Luyster, Katherine Gotham, Whitney Guthrie, Mia Coffing, Rachel Petrak, Karen Pierce, Somer Bishop, Amy Esler, Vanessa Hus, Rosalind Oti, Jennifer Richler, Susan Risi, and Catherine Lord. The autism diagnostic observation schedule-toddler module: a new module of a standardized diagnostic measure for autism spectrum disorders. *J Autism Dev Disord*, 39(9):1305–1320, Sep 2009.

[393] Laina Freyer, Sonja Nowotschin, Melinda K. Pirity, Antonio Baldini, and Bernice E. Morrow. Conditional and constitutive expression of a tbx1-gfp fusion protein in mice. *BMC Dev Biol*, 13(1):33, Aug 2013.

[394] Kentaro Katahira, Kenta Suzuki, Kazuo Okanoya, and Masato Okada. Complex sequencing rules of birdsong can be explained by simple hidden markov processes. *PLoS One*, 6(9):e24516, 2011.

[395] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[396] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[397] Larry Gold, Nebojsa Janjic, Thale Jarvis, Dan Schneider, Jeffrey J. Walker, Sheri K. Wilcox, and Dom Zichi. Aptamers and the rna world, past and present. *Cold Spring Harb Perspect Biol*, 4(3), Mar 2012.

[398] V. A. Shiva Ayyadurai. *Future Visions on Biomedicine and Bioinformatics 1*, chapter Services-Based Systems Architecture for Modeling the Whole Cell: A Distributed Collaborative Engineering Systems Approach, pages 115–168. Springer Berlin Heidelberg, 2011.

[399] Jonathan R. Karr, Jayodita C. Sanghvi, Derek N. Macklin, Miriam V. Gutschow, Jared M. Jacobs, Benjamin Bolival, Jr, Nacyra Assad-Garcia, John I. Glass, and Markus W. Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401, Jul 2012.