ON COMPUTATIONAL STRATEGIES FOR REGULATORY ELEMENT AND REGULATORY POLYMORPHISM DETECTION

by

STEPHEN BLAIR MONTGOMERY

B.A.Sc., The University of British Columbia, 2002

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Genetics)

THE UNIVERSITY OF BRITISH COLUMBIA

September 2006

Abstract

Identification of the mechanisms by which genes are regulated in eukaryotes is one of the principal challenges of modern biology. The emergence of genome sequencing has facilitated the marked expansion of experimental and computational approaches designed to address this challenge. Integrating and assessing this information remains a major scientific endeavor that requires new and innovative application of technology. Furthermore, our limited understanding of the mechanisms of gene regulation in eukaryotes has undermined our ability to understand the role of genetics in gene regulation. Regulatory variants are thought to be responsible for a considerable amount of the heterogeneity within our population and to be fundamental determinants of health. New experimental approaches offer the opportunity to effectively identify markers of disease susceptibility in gene regulatory regions but the discovery of the molecular mechanism of dysregulation remains difficult and time-consuming. It is here where computational approaches are required to prioritize candidate regulatory variants. To do so requires the development of an extensive control set from which characteristic signals can be identified.

This thesis introduces novel approaches for discovering, utilizing, comparing and visualizing regulatory element predictions in completed genomes. This thesis also introduces novel bioinformatics infrastructure for curating regulatory element and variant datasets, and introduces the largest-available, open-access dataset of functional regulatory variants hand-curated from literature. This dataset is used to identify signals which discriminate functional variants from other variants in the promoter regions of human genes using regulatory and population genetics-based computational approaches.

Table of Contents

Abstract ii
Table of Contents iii
List of Tables vi
List of Figures vii
List of Abbreviations viii
Acknowledgements ix
Chapter 1: Introduction 1
1.1 Background 2
1.1.1 Genome browsers 4
1.2 Thesis overview 5
1.3 On bioinformatic analysis of gene regulation 7
1.3.1 Background 7
1.3.1.1 Discovering the cis-acting model of gene regulation 7
1.3.1.2 Current model of gene regulation 8
1.3.1.3 Current strategies for improving our understanding of gene regulation 13
1.3.2 Biological assays for the identification of regulatory elements 13
1.3.2.1 Mutagenesis, reporter genes and binding assays 13
1.3.2.2 High-throughput assays 15
1.3.3 Computational assays for the prediction of regulatory elements 17
1.3.3.1 Signal-based approaches (de novo) 18
1.3.3.2 Database-driven approaches 22
1.3.3.3 Phylogenetic footprinting (Conservation-based approaches) 28
1.3.3.3.1 Conserved non-coding sequences 29
1.3.3.3.2 Transcription factor binding site detection 31
1.3.3.3.3 Phylogenetic shadowing 32
1.3.3.4 Gene function-driven approaches (Expression, ontologies and localization) 33
1.3.3.4.1 Using expression data to identify regulatory elements 34
1.3.3.4.2 Using ontological data to identify regulatory elements 35
1.3.3.4.3 Using localization data to identify regulatory elements 36
1.3.3.4.4 Other data sources 36
1.3.3.5 Architectural features of regulatory regions 37
1.3.3.5.1 Combinatorial (composite) binding of transcription factors 37
1.3.3.5.2 DNA structure and gene regulation 39
1.3.3.5.3 Repetitive elements and gene regulation 39
1.3.3.6 Aggregate approaches (Workbenches and pipelines for regulatory element analysis) 40
1.3.3.6.1 Tools for integrated regulatory analysis 40
1.3.3.6.2 High-throughput regulatory element prediction pipelines 41
1.3.3.7 Performance assessment 42
1.3.4 Synopsis: Future trends 45
1.4 On bioinformatic analysis of genetic mutation 46
1.4.1 Background 46
1.4.1.1 Discovering genetic mutation 46
1.4.1.2 Current model of genetic mutation 48

1.4.1.3 Current strategies for improving our understanding of genetic mutation. 50
1.4.2 Biological assays for the identification and characterization of genetic mutations ... 51
1.4.2.1 SNP Discovery ... 51
1.4.2.2 SNP Genotyping ... 52
1.4.3 Computational assays for the identification and characterization of genetic mutations ... 53
1.4.3.1 Computational approaches for SNP identification ... 53
1.4.3.2 Genetic mutation databases ... 54
1.4.3.3 Computational approaches for SNP characterization ... 57
1.5 On bioinformatics analysis of regulatory mutation ... 60
1.5.1 Background ... 60
1.5.2 Experimental identification and characterization of regulatory SNPs ... 60
1.5.3 Computational characterization of regulatory SNPs ... 62
1.5.4 Synopsis: Future trends ... 65
1.6 Thesis objectives and chapter summaries ... 65
Chapter 2: Sockeye: A 3D Environment for Comparative Genomics ... 71
2.1 Introduction ... 72
2.2 Methods ... 76
2.3 Results ... 78
2.3.1 Sockeye Design ... 78
2.3.2 Data Structure - Storing and managing retrieved annotations ... 79
2.3.3 Sockeye - A GUI Perspective ... 79
2.3.4 3D Viewport ... 81
2.3.5 Integrated Support and Maintenance ... 82
2.4 Discussion ... 82
2.5 Conclusions ... 87
Chapter 3: An application of peer-to-peer technology to the discovery, use and assessment of bioinformatics programs ... 89
3.1 Introduction ... 90
3.2 Methods ... 93
3.2.1 Chinook Architecture ... 93
3.2.2 Client-server communication ... 96
3.2.3 Availability ... 97
3.3 Results and Discussion ... 98
3.3.1 Peer-to-peer for end users ... 98
3.3.2 Peer-to-peer for bioinformaticians ... 100
3.3.2 Application of peer-to-peer approach to assessment of computational tools for transcription factor binding site discovery ... 103
3.4 Conclusions ... 105
Chapter 4: ORegAnno: An open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. ... 106
4.1 Introduction ... 107
4.2 Description of the ORegAnno database ... 108
4.3 Current content of the ORegAnno database ... 112
4.4 ORegAnno Publication Queue ... 113

4.5 Access 114
4.6 Conclusions 114
Chapter 5: A computational discrimination strategy for regulatory polymorphisms in the promoter regions of Homo sapiens. 117
5.1 Introduction 118
5.2 Methods 121
5.2.1 Data 121
5.2.2 Investigated Properties 124
5.2.3 Test design (ALL and GROUP) 128
5.2.4 Support Vector Machine 129
5.2.5 Performance Measurement 129
5.3 Results 131
5.3.1 Discriminant classification 131
5.3.2 SVM cross-validation 134
5.3.3 Distance analysis 135
5.3.4 Surveying new polymorphisms 137
5.3.5 Availability 138
5.4 Discussion 138
5.5 Conclusions 145
Chapter 6: Summary and Conclusions 147
6.1 Summary 148
6.2 Predicting regulatory elements 149
6.3 Prediction regulatory polymorphisms 152
6.4 Conclusions 154
References 155

List of Tables

Table 1: Selected motif discovery algorithms 20
Table 2: Gene regulation databases 22
Table 3: Multiple species conservation scores..... 30
Table 4: Genome-wide regulatory element prediction pipelines for eukaryotes. 42
Table 5: Performance metrics for assessing regulatory prediction tools. 44
Table 6: Gene mutation databases. 55
Table 7: Integrated Bioinformatics Projects 91
Table 8: Bioinformatics tools in Chinook. 97
Table 9: Evidence types and sub-types 110
Table 10: Current contents of the ORegAnno database. 113

List of Figures

Figure 1: Gene transcription machinery 10
Figure 2: PFM Assembly and Sequence Logo. 28
Figure 3: Multiple alignment visualization in Sockeye. 75
Figure 4: Sockeye GUI layout. 80
Figure 5: Comparison of Muscle Specific Regulatory Modules in CACNL1AS/Cacna1s in Human and Mouse. 84
Figure 6: SARs-CoV analysis. 85
Figure 7: Chinook platform. 94
Figure 8: Open Regulatory Competition Correlation Coefficient Table 104
Figure 9: Supporting evidence for regulatory SNP identification (by experiment)..... 122
Figure 10: Supporting evidence for regulatory SNP identification (by record) 123
Figure 11: CpG island positional bias. 131
Figure 12: ALL analysis of 109-rSNPs against ufSNPs. 133
Figure 13: GROUP analysis of 109-rSNPs against ufSNPs. 134
Figure 14: Receiver operating characteristic (ROC) curve for discriminating known regulatory SNPs from polymorphisms of unknown function. 135
Figure 15: Equivalent distance distribution GROUP analysis of 16 rSNPs against 21 ufSNPs. 136
Figure 16: Histogram of positional bias of rSNPs for the first 300bp of sequence 137
Figure 17: Mammalian alignment of OREG0000405 141

List of Abbreviations

CV-CD	Common Variant-Common Disease
EMSA	Electrophoretic Mobility Shift Assay
eQTL	Expression Quantitative Trait Loci
EST	Expressed Sequence Tag
GO	Gene Ontology
INDELS	Insertion/Deletion Polymorphisms
NCBI	National Centre for Biotechnology Information
rSNP	Regulatory Single Nucleotide Polymorphisms
SNP	Single Nucleotide Polymorphisms
SVM	Support Vector Machine
TSS	Transcription Start Site
UCSC	University of California, Santa Cruz
ufSNP	Unknown-function single nucleotide polymorphism
XML	Extensible Markup Language

Acknowledgements

I would like to thank my graduate supervisor Dr. Steven Jones for his support and guidance during my studies—he provided me with many exciting and challenging opportunities which made the Canada's Michael Smith Genome Sciences Centre a very interesting place to work. I am particularly grateful for the educational opportunities he provided as I imagine not many people get to spend the first day of their graduate work at an international genomics conference; these outstanding opportunities to learn throughout my studies have truly made it a unique and rewarding experience. I'm also particularly grateful for his patience in providing me an opportunity to learn academic culture as it pertains to the biological sciences. I am grateful to the Michael Smith Foundation for Health Research and the Natural Sciences and Engineering Research Council of Canada for their salary and research support. I would like to also thank many other people for their tutelage and the time they have taken to work with me, particularly Dr. Wyeth Wasserman, who really helped me find my feet at the beginning of my graduate studies; Dr. Angela Brooks-Wilson, for keeping the informatics portion of this thesis in check; and Dr. Francis Ouellette, who has been instrumental in introducing me to the research community and bioinformatics culture at large—you have all had tremendous faith in my ability, for which I am truly grateful. I would also like to thank Dr. Aly Karsan, for introducing me to the topic of gene regulation and the Genome Sciences Centre, and Dr. Hugh Brock, for believing an engineer could be a geneticist and helping me to enter graduate school. I have enjoyed the friendship and help of many fellow graduate students, among them: Obi Griffith, Monica Sleumer, Erin Pleasance, Johanna Schinas and Ben Good. Special thanks are owed to Mikhail Bilenky and Gordon Robertson for

their long-standing support as part of the GENEREG team and the many other members who have graced this team over its existence. I am truly honored to have worked with so many outstanding people. Thanks are also due to Kai Johnson for editing this manuscript. Finally, I would like to thank my parents, Jane and Zender for their love and constant encouragement and my amazing wife Christine who is the most wonderful and inspirational person I have ever met.

P.S. To my granddad, whose footsteps I am now following, but who never got the opportunity to see it, I will always remember, *"Illegitimus non carborundum."*

Chapter 1: Introduction

1.1 Background

Inheritance is a palate from which a landscape of diversity is created. Each selected brush stroke depicts a lineage that traces back to life's first origins. Each period of this lineage possesses its own style as Van Gogh is to the Cambrian Era what Monet might be to the Silurian Era**.** The study of genetics is an inquisition into the art of life. Its study analyzes how form relates to function, specifically, how change can delineate the boundaries between species and the advantages and disadvantages possessed within a population.

It was Mendel's original discovery that traits are inherited by discrete "hereditary factors," later termed genes, which initiated the study of genetics as a scientific discipline [1]. However, identifying the molecular mechanism of heredity would take nearly a century after Mendel's original discovery. Principally, the relative role of proteins versus nucleic acids as being the primary molecular media for inheritance would remain debated for the early part of the 20^{th} century. Evidence suggesting the primary role of nucleic acids in heredity was obtained in the 1940s by transformations in *Streptococcus pneumoniae;* it was demonstrated that the removal of proteins by proteases allowed the transmission of genetic characteristics, whereas removal of nucleic acids by a deoxyribonuclease enzyme would impede transmission [2]. The principal role of nucleic acids as the molecule of heredity would be definitely proven by Hershey and Chase through differential localization of radioactive proteins and nucleic acids in bacterium and phage coats during viral reproduction [3]. The structure of that nucleic acid molecule, DNA, would be elucidated a year later by Watson and Crick through formative information from Wilkins and Franklin [4,5].

The genetic equivalent of Thomas Wolfe's "kernel of eternity," however, has been the gene, an essential molecular pattern from which all other things proceed[1]. It was in parallel to the discovery of DNA, as the principal molecule of heredity, when insight into the molecular nature of genes started advancing. The locations of genes were determined to be within chromosomes by analysis of sex-linked inheritance patterns of a white-eye mutation in *Drosophila melanogaster* [6]. In 1941, Beadle and Tatum demonstrated that a mutation in a gene location disrupts the process of a metabolic pathway--a formative step towards our modern understanding that genes encode proteins [7]. In 1966, the elucidation of the genetic code that maps amino acids, the basic structural units of proteins, to DNA provided the fundamental link between genes and proteins (reviewed in [8]). As the result of these discoveries, geneticists have since been striving to determine how many genes there are, how each works, and what each is responsible for.

The recent introduction of genomic science has revolutionized research on questions pertaining to the nature of genes and genetics. Modern genomic science was born through the introduction of DNA sequencing in 1977 when Frederic Sanger sequenced the 5368 base pair viral genome of bacteriophage φX174 [9]. For the first time, the full DNA sequence, or genome, of an organism could be determined. Each genome sequenced provided the necessary foundation to begin to identify the full gene complement of an organism and to determine organizational properties relevant to each

[1] Thomas Wolfe (1900-38) was an American short story writer. This quotation is from his work "Of Time and the River". In its original form it is written, "At that instant he saw, in one blaze of light, an image of unutterable conviction, the reason why the artist works and lives and has his being--the reward he seeks--the only reward he really cares about, without which there is nothing. It is to snare the spirits of mankind in nets of magic, to make his life prevail through his creation, to wreak the vision of his life, the rude and painful substance of his own experience, into the congruence of blazing and enchanted images that are themselves the core of life, the essential pattern whence all other things proceed, the kernel of eternity".

gene and its activity. DNA sequencing has since increased at a rate comparable to "Moore's Law"[2] and has fostered the growth of sequence databases in Europe, the United States, and Japan, which now contain over 100 billion base pairs (Aug 2005). In 2001, a 3 billion base pair draft of the human genome was made publicly-available [10]**.** In the 5 years since, the genomes of over 300 organisms have been reported[3], [13] including mammals such as the mouse [14]**,** rat [15], chimpanzee, [16] and dog [17]. This data explosion has created a wealth of challenges as researchers attempt to organize and analyze the features and relationships of many genes and genomes. It has also been a major impetus for the introduction of computers to the repertoire of tools for genetics-based inquiry and the establishment of bioinformatics as a research science.

1.1.1 Genome browsers

The organization of the data around genomes is predominantly achieved through "genome browsers". Most are sophisticated bioinformatics frameworks visualized across the Internet. Among their more basic features, genome browsers, allow a researcher to hone in on particular genomic loci to view the locations of annotations such as genes, conserved elements, repetitive elements, and variants. Depending on which sequenced organism or desired information is required, the choice of genome browser can be quite different; for instance, nematode genome annotation is principally organized by

[2] "Moore's Law" popularly hypothesizes that transistor density will roughly continue to double every two years. Less of a law than posit, it was originally proposed by Gordon Moore, a co-founder of Intel, in 1965 and has since held the test of time.

[3] Completion of a genome is widely-variable depending on genome size and composition. Typically several fold coverage of a genome is required to achieve a gold standard accuracy of approximately 1 error every 1000 bases (reviewed in 11. Chan EY (2005) Advances in sequencing technology. Mutat Res 573: 13-40.). Furthermore, many repetitive, segmentally duplicated, centromeric and telomeric regions remain unsequenced in the human genome due to technological challenges 12. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931-945..

WormBase[18], fruitfly genome annotation by FlyBase[19] and mammalian genome annotation by both the UCSC Genome Browser [20] and EnsEMBL [21]. The typically broad extent of accessibility and analysis provided by these genome browsers and their relative strengths and weaknesses, are outside the scope of this thesis and have been recently reviewed elsewhere [22,23]. Despite their utility, however, genome browsers fundamentally lack the power to display features that may be of importance to the dynamic metabolic processes of the cell (a compelling review of such topics can be found in Webb et al. [24])

1.2 Thesis overview

The goal of this thesis was to develop predictive methodologies for identifying the mechanisms involved in regulation of genes and the impact of single nucleic acid changes upon them with the general aim of improving our understanding of how gene activity is coordinated and how a change in this coordination can interfere with an organism's fitness. Currently, there is a paucity of genes with defined regulatory architecture and scarcer information on the phenotypic effects of mutation on regulatory elements. The reason for this is primarily due to the absence of an accurate, low-cost, high-throughput experimental assay. However, the rapid improvement in sequencing technology and the completion of several model organism genomes, including our own, has facilitated the use of bioinformatics-based techniques to aid in predicting regulatory elements. Moreover, the completion of our own genome, and the application of efficient, low-cost genotyping assays to different human populations, has produced a compendium of common mutations. Both resources have allowed us to use inter- and intra-species

comparisons to characterize the putative regulatory architecture of genes with the benefit of reducing the quantity of targets required for functional validation either in establishing the *de facto* role of a regulatory element or in elucidating the associated phenotype(s) of regulatory mutations. It is known, and will be further illustrated, however, that selection is not sufficient as the growing number of documented functional regulatory mutations have demonstrated that a regulatory element is generally not so highly conserved as to disallow all mutation but not so poorly-conserved as to not matter at all. For this reason, as part of my thesis I have focused on combining information derived from population studies with comparative analyses and information indicative of regulatory potential.

My overall approach to this thesis has been in three parts: the first part has been to improve assessment and interpretation of regulatory element detection methodologies, the second part has been to categorize documented regulatory elements and functional regulatory mutations to improve the utilization of this knowledge in genomic sciences, and the third part has been the development of a strategy for predicting functional regulatory mutations by classifying different genomic features associated with regulatory potential and natural selection. This comprehensive approach takes principal advantage of current research in the bioinformatics of gene regulation and population genetics. The majority of this introduction will provide the necessary background into these two active fields of research that has guided the research I have conducted as part of this thesis.

1.3 On bioinformatic analysis of gene regulation

1.3.1 Background

The study of gene regulation aims to identify the processes which control gene expression. This control is intrinsically responsible for an organism's development and upkeep and ultimately describes fundamental aspects of its evolutionary ascendance.

1.3.1.1 Discovering the cis-acting model of gene regulation

The first breakthrough in understanding how genes are regulated at a molecular level was through characterization of the regulatory dynamics of lactose catabolism in the *lac* operon of *Escherichia coli* [25]. When the *lac* operon is expressed, two lactose metabolism genes are expressed, *lacY* and *lacZ*. The protein products of each respectively are *lactose permease,* which facilitates transport of lactose through the cell membrane, and *B-galactosidase,* which converts the disaccharide lactose into the monosaccharides, glucose, and galactose. It had been originally observed, however, that *E. coli* was able to preferentially metabolize glucose and lactose depending on each metabolite's presence in growth media ([26]; reviewed in [27]). This preferential metabolism could be observed using *X-gal* (a lactose analog, which is typically colorless except blue when it is hydrolyzed by *B-galactosidase)*. *E. coli* cultures in media consisting solely of glucose and *X-gal* would remain white while those in lactose and *X-gal* media would turn blue. Furthermore, when *E. coli* were irradiated, some of the cultures in glucose media would also turn blue. By mapping this mutation to a new gene location, called *lacI*, Jacob and Monod hypothesized that this new gene product acted to repress the expression of the *lac* operon. To confirm this, *E. coli* were transfected with

an extra copy of the *lacI* gene and irradiated. Considering that it was a lower probability that *lacZ* expressing cells in glucose media were the product of two *lacI* mutations, identified mutants were theorized to have mutations in the regulatory control region of the *lac* operon. These mutant sequences, when transfected into wild-type *E. coli,* would still produce the mutant phenotype suggesting that the mutation is *cis*-acting in that the wild-type non-coding region is bound by *lacI* to repress lactose catabolism.

These observations remain of fundamental importance to our current understanding of gene regulation. The role of proteins (termed transcription factors, which bind control sequences in DNA to alter the expression activity of adjacent genes) has only become more predominant. It is the comprehensive identification of both the transcription factors and their cognate binding sites that remain a primary focus for many biologists.

1.3.1.2 Current model of gene regulation

Our current understanding of gene regulation as it applies to higher-order eukaryotes encompasses a variety of architectural components which aid in facilitating gene expression. The central components for each protein-coding gene typically include at least one proximal promoter region controlling the constitutive expression of the target gene and multiple enhancers or locus-control regions governing tissue- and stage-specific expression of multiple genes[4] (see Figure 1). Characterization of well-known regulatory

[4] Several longer-range and more generic mechanisms exist for controlling gene expression but will not be covered here including the repressive effects of small ncRNAs, transcript stability modifiers, boundary elements (insulators) which control the extent of a regulatory regions effect and histone acetylation and DNA methylation status which both effect the accessibility of transcription factors to regulatory regions (reviewed in 28. Ross J (1996) Control of messenger RNA stability in higher eukaryotes. Trends Genet 12: 171-175, 29. Almeida R, Allshire RC (2005) RNA silencing and genome regulation. Trends Cell Biol 15: 251-258, 30. Robertson KD (2005) DNA methylation and human disease. Nat Rev Genet 6: 597-610, 31.

regions (promoters and enhancers) has demonstrated that there are typically four to eight transcription factor binding sites per region [38]. The composition of each regulatory region encodes a specific program for the development and maintenance of an organism; each attenuates expression levels as necessary in the formation of different morphologies and in response to external stimuli.

Chen ZX, Riggs AD (2005) Maintenance and regulation of DNA methylation patterns in mammals. Biochem Cell Biol 83: 438-448, 32. Holmes R, Soloway PD (2006) Regulation of imprinted DNA methylation. Cytogenet Genome Res 113: 122-129, 33. Attwood JT, Yung RL, Richardson BC (2002) DNA methylation and the regulation of gene transcription. Cell Mol Life Sci 59: 241-257, 34. Mizzen CA, Allis CD (1998) Linking histone acetylation to transcriptional regulation. Cell Mol Life Sci 54: 6-20, 35. Bashirullah A, Cooperstock RL, Lipshitz HD (2001) Spatial and temporal control of RNA stability. Proc Natl Acad Sci U S A 98: 7025-7028, 36. Capelson M, Corces VG (2004) Boundary elements and nuclear organization. Biol Cell 96: 617-629, 37. Udvardy A (1999) Dividing the empire: boundary chromatin elements delimit the territory of enhancers. Embo J 18: 1-8.).

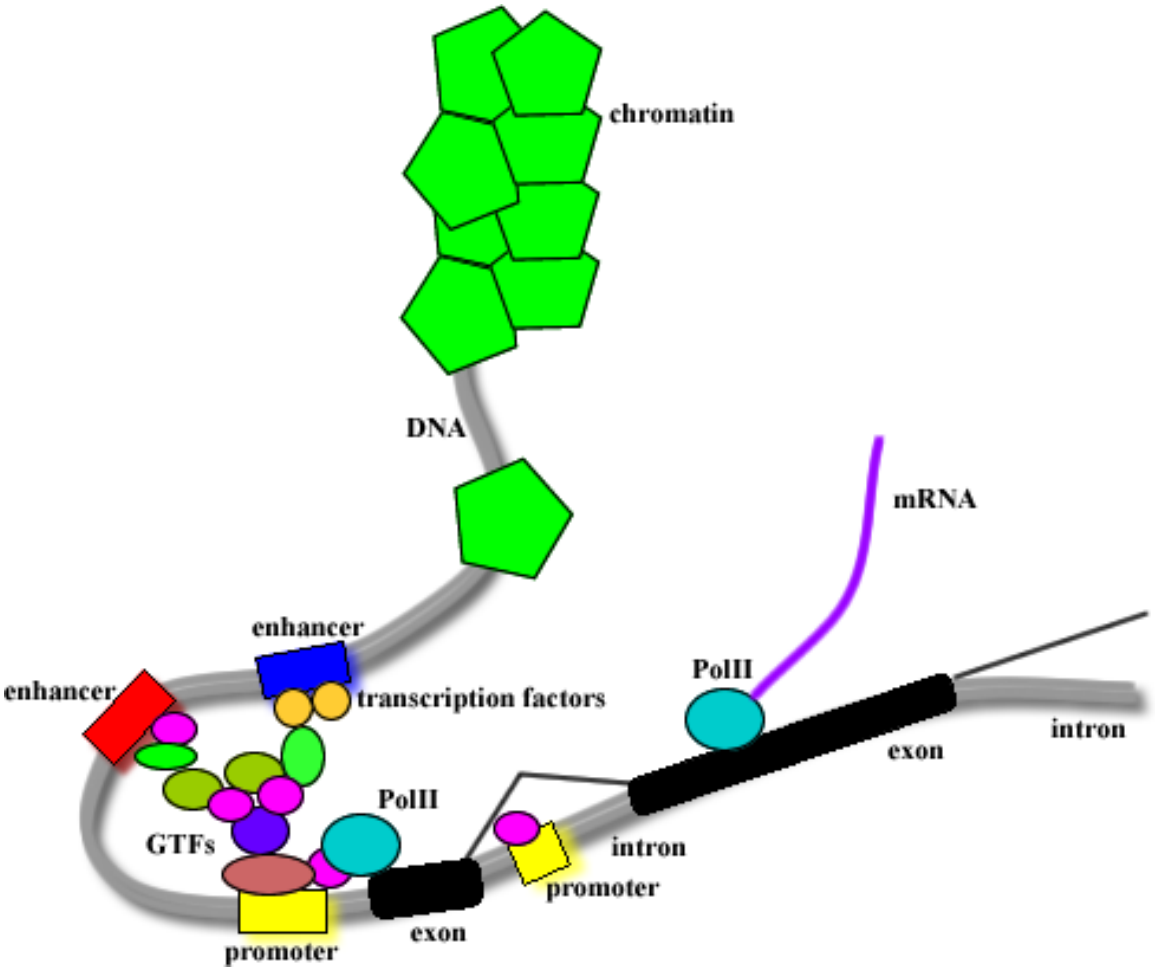

Figure 1: Gene transcription machinery

Chromatin (green) is uncondensed to allow gene transcription. General transcription factors bind promoter regions (yellow) to initiate basal transcription. Activators or repressors (transcription factors) bind enhancers (red and blue) to modulate transcription in a tissue- and stage-specific manner. RNA polymerase II (blue) transcribes messenger RNA (purple) as it traverses the gene's DNA (grey).

The role of a promoter region is typically to initiate transcription by positioning of the core transcription machinery. The upstream non-coding region immediately proximal to a gene has canonically been the first region investigated for "regulatory potential", or its ability to drive gene expression, when identifying a promoter. These candidate promoter regions are typically easy to find as extensive cDNA and expressed sequence

tag (EST) libraries can be used when available to demarcate the coding boundaries of a particular gene and, in cases where they do not provide sufficient coverage, gene identification can be supplemented or complemented with gene prediction software. However, identifying the locations of alternative promoters, which can regulate the production of alternative transcripts, has required definitive determination of the locations of transcription start sites (reviewed in [39]). Recent genome-wide discovery of transcription start sites in human and mouse, has also found supporting evidence for promoters existing within exons [40] or within the 3' end of genes [40-42]. Inspection of candidate promoter regions proximal to the transcription start site has identified two classes: conserved promoters containing a TATA-box and an INR (initiator) element, and promoters, which are less conserved but more CpG rich [40,43]. The former class has been studied in detail and generally involves the recruitment of a combination of general transcription factors (GTFs) to generate a pre-initiation complex, which initiates RNA polymerase II-mediated transcription (reviewed in [44,45]). The latter class is found predominantly in the promoter regions of house-keeping genes. Since Methyl-C of CpG dinucleotides tends to decay to TpG/CpA dinucleotides, their presence suggests functional constraint in these regions [46]. Furthermore, the presence of CpG islands has been identified to help maintain the methylation status of associated genes and it has been proposed that they are associated with promoters that are transcriptionally active in totipotent stages of development [47].

The role of enhancers and locus control regions is generally to facilitate spatio-temporal transcriptional activity at physiological levels for a set of genes. Their function is not necessarily too different from that of promoters, as RNA polymerase II and GTFs

have been identified as being recruited to enhancer elements [48-51]. This has suggested that enhancers act as centres for transcription factor recruitment [52], and further suggests their role in facilitating RNA-Polymerase II-mediated transcription in CpG promoters. Enhancers are generally different from promoters, in that they are typically orientation-independent, can be located distally to their target genes, and that their transcriptional control program can modify the basal level of gene expression in a copy number-dependent manner (reviewed in [53-55]). Only a few enhancers and locus control regions have been extensively researched; among the best examples of these are the *ß-globin* LCR in human [54] and the *Evenskipped stripe* 2 enhancer in *D. melanogaster* [56].

Within either enhancers or promoters, specific combinations of transcription factors bind to target DNA sites to execute a transcriptional program. As discussed, a typical enhancer or promoter contains 4-8 transcription factor binding sites and both can recruit GTFs and RNA polymerase II. There is also a large variety of other transcription factors that can be sequestered to individual regulatory regions. In humans, there are an estimated 1900 transcription factors which can be organized into individual transcription factor families based on DNA binding domains including zinc finger proteins, helix-turn-helix proteins, leucine zipper proteins, and helix-loop-helix proteins [57]. Each binds a characteristic DNA motif tolerant of degeneracy that can stretch from as small as 5bp for Pax-4 or HOXA5 to as long as 30bp for HSF or OCSBF-1. The small size of transcription factor binding sites, the quantity of available transcription factors, and the lack of precise, high-throughput identification techniques have made the determination of these elements and associated transcription factors analogous to finding the proverbial needle in a haystack.

1.3.1.3 Current strategies for improving our understanding of gene regulation

Defining which transcription factors are involved in a particular transcriptional program of an individual gene requires considerable effort, let alone identifying the complete regulatory network (all possible transcriptional programs) for an organism. An international consortium has initiated the ENCODE project with the aim of identifying all the functional elements in 1% of the human genome (30 Mb) [58]. Their approach has required a combination of molecular biology-based and computational assays. These types of strategies will be discussed in the remainder of section 1.3 with an extended focus on the developments in computation-based techniques.

1.3.2 Biological assays for the identification of regulatory elements

1.3.2.1 Mutagenesis, reporter genes and binding assays

Classical genetics-based detection of regulatory regions typically involves finding a measurable mutant phenotype through *trans*-acting or *cis*-acting mutation and/or extracting a minimal amount of DNA to drive the expression of a reporter gene (typically luciferase, chloramphenicol acetyl transferase (CAT), green fluorescent protein (GFP), or *lacZ*.) Jacob and Monod were able to take advantage of the former technique by identifying equivalence between the phenotype of mutation in the *lac* promoter region and mutation in the *lacI* transcription factor. The latter technique, however, has been more systematically employed in techniques such as "promoter-bashing" or "linker scanning." In "promoter-bashing," a series of truncated promoter fragments are fused to a reporter gene and the relative effects on reporter gene expression activity are compared. This technique, however, lacks the ability to detect expression driven by combinations of

transcription factors. Although it is a more time-consuming and expensive experiment, this limitation can be overcome in "linker scanning," where a promoter fragment, fused to a reported gene, is selectively mutated and the relative effects on reporter gene expression activity are compared.

The detection of transcription factor binding sites has traditionally been performed using electrophoretic mobility shift assays (EMSAs) or chromatin immunoprecipitation. In EMSAs, DNA fragments are tested for their protein-binding ability through observation of normal or impeded (protein-bound) gel shift patterns. The implicated transcription factor(s) can be identified using a variant of this technique called a supershift; DNA-protein complexes are incubated with antibodies against the transcription factor of interest and when the correct antibody is found a further retardation of the gel shifting pattern is observed. In chromatin immunoprecipitation, protein and DNA are cross-linked using formaldehyde then sheared into smaller regions and precipitated using antibodies against the transcription factor of interest. The cross-links are then reversed and DNA can be purified from the precipitate and amplified to determine if it is from the region of interest. These procedures are generally considered laborious as without *a priori* information regarding the DNA fragments and transcription factors to test, the test parameter space expands exponentially.

The limitations of all these techniques are not only that they are generally difficult and time-consuming, but they are limited to solely identifying the regulatory regions and transcription factors involved in the physiological conditions tested. Systematic effort to characterize regulatory elements in multiple cell lines are currently underway as part of

the ENCODE project; analysis of promoters predicted for transcription start sites in 16 different cell lines has identified 387 fragments that function as promoters [59].

1.3.2.2 High-throughput assays

Whole genome sequencing has facilitated the development and application of high-throughput assays for the identification of regulatory elements. Two principal technologies are microarray-based chromatin immunoprecipitation (ChIP-chip) and DNaseI hypersensitivity site mapping.

The ChIP-chip assay was originally applied to the yeast genome to identify binding sites for the transcription factors Gal4 and Ste12 [60]. In this study, chromatin immunoprecipitation was initially performed and then both the immunoprecipitated DNA and the unenriched DNA were differentially labeled with fluorescent dyes. Each labeled DNA population was hybridized to a microarray spotted with 6361 intergenic regions. Fluroescence intensity was measured to identify intergenic sequences corresponding to the enriched DNA fragments. Ten genes were identified as having binding sites for Gal4 and twenty-nine genes for Ste12. A principal challenge with this technique has been that it is dependent on the number of regions that can be spotted to a microarray and the size of the sheared fragments; mapping of transcription factor binding sites can only be resolved to a resolution of 1-2kb. This technology has since been combined with high-density oligonucleotide arrays to map transcription factor binding across ENCODE regions for E2F1, MyC and RNA Polymerase II [61], human chromosomes 21 and 22 for the estrogen receptor, Sp1, cMyc, and p53 [62,63], human chromosome 22 for CREB [64], selected regions for menin, MLL1 and Rbbp5 [65] and 25% of the human genome

for the general transcription factor TBP1 [66]. The majority of these studies suggest that many thousands of transcription factor binding sites exist for each tested transcription factor; this currently presents a further challenge to this technology as the promiscuity of binding sites raises speculation as to the extent that binding is direct and/or will indicate function [67,68]. A promising advance in ChIP-based technology has been described by coupling chromatin immunoprecipitation with paired-end ditag (PET) sequencing [69]. ChIP-PET has been able to extend transcription factor binding site identification to the whole genome while narrowing the target binding site to less than 100 bp; however, the cost for each experiment still remains considerable.

DNaseI hypersensitivity site mapping is performed by digesting DNA with deoxyribonuclease 1 (DNaseI) to cleave protein-free DNA. Cleaved DNA is hybridized with a radio-labeled probe for a target region of interest. The sizes of the hybridized fragments are used to map sites that are devoid of protein-binding. This technique is considered laborious and has traditionally been used in single genes. Genomic library-based techniques, however, are now being applied to determine DNase hypersensitive sites across multiple loci [70-72]. But like ChIP-chip experiments, these assays have yet to be sufficiently scaled to a capacity that will efficiently survey a mammalian genome, let alone mammalian cell lines.

A future highlight of these types of approaches will be their use in comparing binding occupancy under different physiological conditions.

1.3.3 Computational assays for the prediction of regulatory elements

A salient challenge with pattern discovery in biological sequences, as used to detect transcription factor binding sites, has been that alignments are often insufficient for identifying short re-occurring patterns. In other words, there is typically no common organization in the biological sequences except for a short pattern which needs to be recognized. Furthermore, the lack of a low-cost, highly-accurate biological assay for identifying the location of regulatory elements in mammalian genomes has advanced the development of a variety of bioinformatics methodologies and many more algorithms with the aim of identifying and prioritizing candidate regulatory elements (reviewed in [73-81]). These methodologies can be broken down into five major classes: those that: 1) use a signal-based approach, where a promoter or transcription factor binding site is determined from sequence composition; 2) use a database-driven approach, where new predictions are made from previously constructed promoters or transcription factor binding models; 3) use comparative genomics, where sequence conservation over multiple organisms implies functional constraint; 4) use function-based information (commonly from coexpression and gene ontology data); and 5) use architectural features of regulatory regions, such as combinatorial binding patterns or DNA curvature. These methodologies are not mutually exclusive as comparative and coexpression-based approaches typically depend on *de novo* motif discovery algorithms or pre-existing transcription factor binding models. This section will give considerable attention to each of these methodologies and will discuss aggregate approaches and assessment techniques.

1.3.3.1 Signal-based approaches (de novo)

Computational detection of regulatory elements is intrinsically dependent on an algorithm's ability to detect a signal that discriminates a functional regulatory element from the null hypothesis. While this is the *modus operandi* of all regulatory element prediction algorithms, a subset of those algorithms attempt to detect these signals from sequence composition alone. This has most commonly manifested itself in algorithms that predict regulatory regions from only GC and CpG content and algorithms that predict transcription factor binding sites by searching a collection of sequences for short words or 'motifs' which are overrepresented.

1.3.3.1.1 Regulatory region prediction from GC and CpG content

GC and CpG-island content are sequence composition metrics that have a known relationship to regulatory regions and the maintenance of gene expression in mammals [82-84]. Both low GC and CpG island content have a known correlation with chromatin-mediated suppression [85]. In vertebrates, CpG dinucleotides are frequently methylated and often mutate to TpG dinucleotides, except when found in high concentrations in CpG islands where they remain unmethylated (reviewed in [86-88]). In human genetic diseases and cancers, CpG islands are frequent targets of mutation, highlighting their importance in maintaining normal gene expression [89]. Furthermore, classes of promoters are distinguishable from their GC composition around the transcription start sites [82,90]. These distinguishing properties have been exploited to identify promoter sequences in computational tools, like Eponine [91], CpGProD [92], FirstEF [93], and CpGpromoter [94]. The relationship of CpG islands and regulatory regions, however, is

not necessarily a direct one and is complicated by the existence of alternative types of regulatory regions ([43]; reviewed in [86]).

1.3.3.1.2 de novo motif discovery

In 1979, the TATA-box was first elucidated by comparison of upstream non-coding sequences from a variety of species [95]. It was the first transcription factor binding site identified in the core promoter of eukaryotes using direct sequence comparison. Since this discovery, the comparison of sets of sequences for short, overrepresented patterns has become the *de facto* method for detecting new and previously uncharacterized transcription factor binding sites in regulatory regions; and the majority of advances have been in the optimization of the input and background sequence sets and the number, variability, and length of patterns identified.

Motif discovery algorithms are designed to aid detection and modeling of short patterns that are overrepresented in a set of sequences. Most are designed to allow a degree of degeneracy which is representative of the relaxed specificity of most transcription factor–DNA binding interactions. Two principal types of motif discovery algorithms are generally available: those that are enumerative and those which are designed to satisfy an objective function implemented based on *a priori* assumptions regarding transcription factor binding. The enumerative algorithms, or "word counting" algorithms, compile exhaustive lists of patterns; they are constrained by sampling space and model assumptions. Specifically, for purely enumerative algorithms, the execution time increases exponentially with motif size. Objective function algorithms for motif finding have typically used modeling parameters to calculate probabilities for nucleotide

arrangements at different positions within a motif, the most popular of which are Gibbs samplers, which are discussed in Section 1.3.3.1.2.1. Other types of probabilities can also be maximized, for example, by assuming there are a fixed number of binding sites per input sequences (most algorithms typically assume there is only one) [96]. Performance of objective function algorithms are challenged by those sequences that do not contain a motif, contain multiple motifs, or have particularly strong motifs masking functional but more degenerate motifs. Both objective function and enumerative algorithms are challenged by transcription factor binding sites that do not satisfy predefined motif width constraints.

Table 1 describes several well-used motif discovery programs which have been investigated further as part of this thesis; for more information, a complementary list has also been published by Tompa et al. [97]. It should be noted that neither list is complete; one recent paper has suggested over a hundred motif discovery tools and provides a thorough breakdown of the algorithms used by these techniques [81].

Table 1: Selected motif discovery algorithms

Program	Operating description
Teiresias [98]	An enumerative method that uses seeded sub-patterns to reconstruct maximally-specific motif patterns.
GLAM [99]	Gibbs sampling-based algorithm that automatically optimizes the alignment width and evaluates the statistical significance of its output.*
ANN-Spec [100]	Models the DNA-binding specificity of a transcription factor using a weight matrix.*
Weeder [101]	Consensus-based method that enumerates exhaustively all the oligos up to a maximum length and collects their occurrences (with substitutions) from input sequences.*
AlignACE [102]	Gibbs sampling algorithm that returns a series of motifs as weight matrices that are overrepresented in the input set.*
MDscan [103]	Designed for ChIP-array-selected sequences. Combines

	word enumeration and position-specific weight matrix updating.
MotifSampler [104]	Matrix-based, motif-finding algorithm that extends Gibbs sampling by modeling the background with a higher order Markov model.*
MEME [105]	Optimizes the E-value of a statistic related to the information content of the motif.*
CONSENSUS [106]	Models motifs using weight matrices, searching for the matrix with maximum information content.*
ELPH [107]	An open source Gibbs sampling algorithm. Assumes that each sequence contains one copy of the motif.
Recursive Gibbs Sampler [108]	A Gibbs-sampling algorithm that incorporates a Bayesian method for inferring the number and the locations of the TFBS for multiple TF motifs simultaneously.

*Operating descriptions as previously reported in Tompa et al. [97].

1.3.3.1.2.1 Gibbs samplers

Gibbs sampling is a statistical method that has become popular in motif discovery algorithms, because of speed and sensitivity. The original incarnation of the Gibbs sampler, as published in 1993, required that a single motif was present in each of the input sequences being analyzed [109]. These sequences are then scanned to maximize the probability of observing a particular pattern against the background probability. Specifically, a two stage approach is conducted:

1) Predictive update step: A sequence is withheld and a randomly-selected pattern description and background frequencies are calculated from the remaining n-1 sequences.

2) Sampling step: Every possible motif of predetermined width within the withheld sequence is considered as a candidate for the pattern description (generated in Step 1). Each motif is weighted according to

its probabilities from which a new pattern position is randomly selected.

Using these steps, once a pattern is found by chance (whether complete or not) subsequent iterations will weigh the pattern discovery as to drive it towards establishing the strongest pattern. Since the inception of this technique, several tools have been developed using and improving Gibbs sampling strategies; specifically, the original authors of the 1993 publication have recently used heuristics approaches to allow the Gibbs sampler to detect multiple conserved patterns [108].

The major difference between Gibbs sampling approaches and Expectation Maximization (EM) approaches is that Gibbs sampling attempts to optimize a probability function using unknown variables were EM approaches attempt to optimize a probability function based on the means and modes of unknown variables through expected statistics.

1.3.3.2 Database-driven approaches

The computational prediction of regulatory regions and transcription factor binding sites has been largely facilitated by the existence of databases describing promoters and transcription factor binding sequences, of which there are several (see Table 2). Each of these types of databases has provided a unique resource for addressing specific challenges with regulatory element detection.

Table 2: Gene regulation databases

The Arabidopsis *cis*-regulatory element database (AtcisDB) [110]	AtcisDB consists of transcription factor binding site information, promoter sequence, and related annotations for *Arabidospis thaliana*. Core promoters are predicted from full-length cDNAs. 25,806 promoters sequences were annotated as of September 2005.

Arabidopsis thaliana Promoter Binding Element Database (AtProbe) [111]	Resource integrating regulatory element information for *Arabidopsis thaliana* from Entrez, PlantCARE, PLACE, PubMed, and TRANSFAC. 172 binding sites for 118 binding elements.
Arabidopsis transcription factor database (AtTFDB) [110]	Contains 1,690 Arabidopsis transcription factors and their sequences (protein and DNA) grouped into 50 (October 2005) families with information on available mutants in the corresponding genes.
C. elegans promoter database (CEPDB) [112]	Contains promoters for 618 *C.elegans* genes as of June 2006.
DBD: Transcription factor prediction database [113]	Contains transcription factor predictions for organisms based on homology through HMM modeling of domains. Consists of predicted transcription factors for 150 completely sequenced genomes (37736 transcription factors).
Drosophila DNase1 Footprint Database [114]	A dataset based on a systematic literature curation and genome annotation of DNaseI footprints for *D. melanogaster*. Contains 1367 binding sites for 87 transcription factors in 101 target genes from 201 primary references.
Eukaryotic Promoter Database (EPD) [115]	The Eukaryotic Promoter Database is an annotated non-redundant collection of eukaryotic POL II promoters, for which the transcription start site has been determined experimentally.
The *Drosophila* transcription factor database (FlyTF) [116]	A database of fruitfly transcription factors. Contains 753 putative site-specific transcription factors of which 454 are well-supported.
Hematopoiesis Promoter Database (HemoPDB) [117]	HemoPDB is composed of experimentally defined regulatory information, including TFs, *cis*-regulatory elements, their target gene promoters and corresponding annotations, with links to supporting published references with respect to hematopoiesis.
JASPAR [118]	An alternative to TRANSFAC, this database is open access. This database produces tightly-controlled binding profiles with strict quality restrictions. This database provides a programming API for improved access.
The Liver Specific Gene Promoter Database (LSPD) [119]	Contains liver-specific promoters and transcription factor binding sites. As of June 2006, 178 specific genes are listed with 368 regulatory elements.
Mammalian Promoter Database [120]	Collection of promoter databases for human, mouse, and rat.
MPromDb [121]	Database for gene promoters with experimentally supported annotation of transcription start sites (TSS), *cis*-regulatory elements, CpG islands, and ChIP-chip experimental results.
Orthologous Mammalian Gene Promoter datababse (OMGProm) [122]	A resource of mammalian gene promoters and their orthologs between humans and rodents.

Osteo-Promoter Database (OPD) [123]	Database that analyzes promoters of genes which differentiate along with the osteogenic pathway.
PLACE [124]	PLACE is a database of motifs found in plant *cis*-acting regulatory DNA elements as annotated from literature. Contains 453 entries as of May 2006.
PlantCARE [125]	PlantCARE is a database of plant *cis*-acting regulatory elements, enhancers, and repressors. Contains 668 *cis*-acting regulatory elements.
PlantProm DB [126]	An annotated, non-redundant collection of proximal promoter sequences for RNA polymerase II with experimentally determined transcription start site(s), TSS, from various plant species.
Promoter Database of *Saccharomyces cerevisiae* (SCPD) [127]	Yeast promoter database containing multiple promoters and transcription factor binding sites.
Regulatory Element Database for *Drosophila melanogaster* (REDfly) [128]	REDfly is a curated collection of known *Drosophila* transcriptional *cis*-regulatory modules (CRMs). Contains 628 regulatory elements as of April 2006.
Riken Transcription Factor Database (TFdb) [129]	Contains non-redundant transcription factors predicted for mouse.
TRANSFAC [130]	A widely-used transcription factor site and matrix database. Curators annotate transcription factor binding sites from literature to assemble representative consensus sequences and position weight matrices. Binding profiles are of diverse quality and possess some redundancies. This database is not open access.
Transcriptional Regulatory Element Database (TRED) [131]	A mammalian regulatory element database. Contains genome-wide predictions of core promoters for human, mouse and rat. Also contains expert curated transcription factor binding sites for cell-cycle factors either computationally or experimentally determined.
Transcriptional Regulatory Regions Database (TRRD) [132]	Contains information on structural and functional organization of transcription regulatory regions of eukaryotic genes. As of April 2006, contained over 10000 transcription factor binding sites and 3490 regulatory regions curated from 7609 references.

Promoter databases facilitate discovery of indicators or discriminants of regulatory potential, such as CpG prevalence [133] or evolutionary conservation [134], and elements of regulatory architecture, such as core promoter composition [135] or

prevalence of promoter polymorphisms [136]. Furthermore, promoter databases facilitate computational approaches to transcription factor binding sites by providing noise or background models in which prediction algorithms can be trained to detect novel signals. For instance, MotifSampler uses the EPD database to generate clade-specific background models online [137].

Transcription factor binding site databases, like TRANSFAC and Jaspar, have been developed to describe the diversity of sequences bound by a single transcription factor. Each database has predominantly benefitted from *in vitro* binding assays, like SELEX experiments (reviewed in [138]), to determine the sequence targets of transcription factors. The utility of these databases is that a transcription factor's ability to bind specific sequences can be modeled and subsequently applied to novel sequences.

Transcription factor binding models within TRANSFAC and Jaspar are represented as either an IUPAC consensus sequence or a position specific weight matrix. IUPAC consensus sequences were originally used to describe mutability between base positions of a transcription factor binding site by representing variant nucleotides with an enriched set of symbols. The disadvantage of this encoding is that the quantitative predisposition of individual bases to promote binding is essentially ignored; weight matrices were introduced to include this information [139]. A weight matrix is assembled by measuring the frequencies of individual nucleotides at each position in a binding site. There are several variants of these types of matrices which only ameliorate types of mathematical manipulation when scanning novel sequences for their amenability to binding. Weight matrices are discussed more in Section 1.3.3.2.1.

The difficulty with application of transcription factor binding models, such as those represented by IUPAC consensus sequences and weight matrices, is that they are prone to making Type II errors (false positives). Conservative application of a weight matrix for the transcription factor MEF2 predicts a binding site every 350bp [140]. It has also been observed that when using the complement of matrices in a TF database, a binding site is generally predicted at every base (Wasserman, WW, personal communication). To describe this relative to results obtained from biological assays, if a typical ChIP-chip result identifies 10,000 binding sites across the human genome for a single transcription factor, a matrix scan could typically yield about 3 million such sites – when unfiltered due to location, suggests that only 1/3 of a percent of all such predictions can possibly be true. The ability of such models to accurately identify true binding sites is clearly limited by the number of transcription factor binding models being used and the size of the genomic sequence being searched. The next three sections will discuss specific methodologies to improve both of these parameters.

1.3.3.2.1 Position-specific Scoring Matrices

Position-specific Scoring Matrices (PSSMs) describing protein-DNA binding specificities have been used to describe the binding sites of individual transcription factors for nearly 25 years ([139]; reviewed in [141]). In their simplest form, they are called Position Frequency Matrices (PFMs), where the observed frequency of a nucleotide is computed at each position of an alignment. An alternative form of PSSM, called a Position Weight Matrices (PWM), is computed by taking the logarithm of the frequencies at each position. This type of matrix is commonly used due to

mathematically tractability since potential binding site scores are calculated by summation across nucleotide positions. A lesser used type of matrix is the Information Content Matrices (ICMs), which describe how different a position is from its expectant distribution. Positions that are perfectly conserved are described as containing 2-bits of information, whereas positions that are shared perfectly between two nucleotides are characterized as having 1-bit of information. The information embedded within a PSSM has typically been displayed as a sequence logo, where the size of a nucleotide at a position represents its frequency or information content [142]. A representative PSSM and sequence logo are displayed in Figure 2.

PSSMs have been widely used, not just for their improved specificity over consensus sequence approaches, but because the matrices have been shown to be concordant with binding strength [143]. Among their disadvantages, however, they have no capabilities for describing more complicated models of binding, including the insertion of variable gaps or dependent relationships between nucleic acid positions. A recent analysis of the latter has suggested that 25% of transcription factor binding motifs show correlations between positions [144]. Further confounding the validity of existing predictions, it is of cautionary note that binding sites that have not been observed *in vivo* can be predicted strongly and recent evidence has suggested that existing PWMs are too broadly defined and can be subclassed further using a mixture model (i.e. there may be multiple PWMs which better classify the set of binding interactions for a particular transcription factor) [145].

For working with PSSMs, the TFBS-Perl package provides a programming interface to databases like TRANSFAC and Jaspar which contain PSSM information,

allow the manipulation of PFMs, PWMs and ICMs and the generation of representative sequence logos [146]. Alternately, sequence logos can also be constructed online using WebLogo [147].

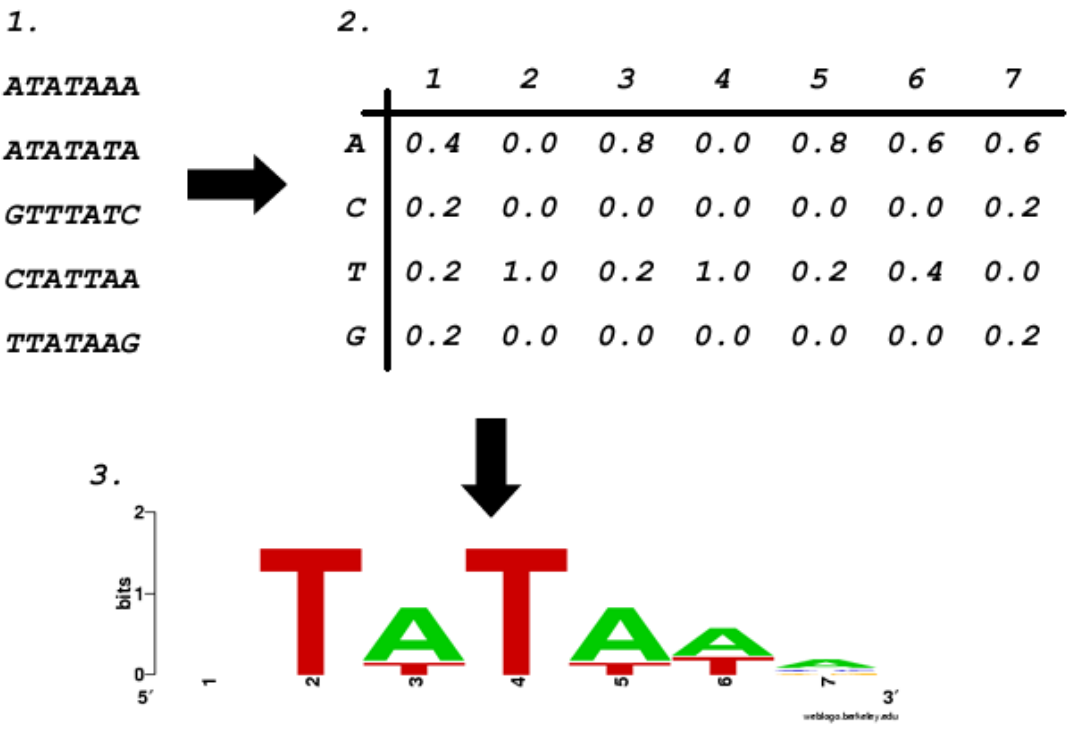

Figure 2: PFM Assembly and Sequence Logo.

1) Identified binding sites are assembled into a set and 2) a PFM is constructed by calculating the frequency of nucleotides at each position in the set. 3) PFMs are commonly represented as sequence logos describing the information content at different positions in the matrix (this sequence logo was constructed using WebLogo [147]).

1.3.3.3 Phylogenetic footprinting (Conservation-based approaches)

Phylogenetic footprinting is a technique useful for detecting regulatory elements from information obtained from cross-species comparisons [148]. The principle of these analyses is to compare the mutation rate of orthologous, non-coding DNA segments in two or more organisms to a neutral mutation rate. Those segments that possess mutation

rates lower than the neutral rate are hypothesized to have been selected due to functional constraint (reviewed in [149]). The utility of this approach has demonstrated its power to predict functional transcription factor binding sites *in vitro* ([150]; reviewed in [151]). Furthermore, the fecundity of genome sequencing projects has facilitated the expansion of phylogenetic footprinting approaches to many different species (reviewed in [152]). This section will discuss how putative regulatory regions are being detected through genome-wide approaches to detecting conserved non-coding sequences and then highlight how these conserved regions are being used for transcription factor binding site identification.

1.3.3.3.1 Conserved non-coding sequences

The utility of conservation-based approaches to regulatory region detection has been highlighted in initial comparative analysis of the human and mouse genome [14]. Comparison of known regulatory regions (n=95) supports a predisposition for regulatory regions to possess mutation rates lower than the neutral rate. Furthermore, it was estimated that, at minimum, 5% of the human genome is under this type of constraint. By identifying these constrained regions, computational techniques have been developed to extract candidate regulatory regions *en masse* ([134,153]; reviewed in [154]).

A major challenge with identifying conserved regions under functional constraint is in selection of an appropriate evolutionary divergence for comparison [155-158]. Each comparison is limited to finding functional elements common to the shared ancestry of the species under observation. Furthermore, when comparing closely related species, their may not have had enough time to accumulate sufficient mutation while highly-

divergent species may have had too much time, thereby permitting adaptive or lineage-specific change. (The evolution of regulatory regions will be discussed in more detail in section 1.5.)

Multiple species sequencing and comparison have aimed to mitigate the challenge of selecting suitable species by delineating selection pressures across different clades of organisms [159]. To facilitate these analyses, computation techniques and resources have been developed to aid in identification and interpretation of multiple sequence comparisons (reviewed in [160]). Several of the current techniques that have been designed explicitly for measuring conservation in multiple species are discussed in Table 3.

Table 3: Multiple species conservation scores

MCS scores [161]	a. *Binomial-based*: Conservation scores are calculated individually for 25bp windows from a multiple-sequence alignment for human and each other species. More distantly diverged species are weighted more heavily. MCS scores are derived by averaging the scores moving up through the associated phylogeny. b. *Parsimony-based*: A parsimony score is derived for each site in a multiple sequence alignment. A neutral substitution model is used to calculate the probability of the parsimony score given the composition of the phylogenetic tree. The final value is averaged over a 25bp window centered on analyzed site.
PhastCons scores [162]	PhastCons scores are derived by a phylogenetic Markov model which is designed to transition through conserved and non-conserved states in a reference genome while comparing the phylogeny at each nucleotide. Predicted elements are assigned a log-odds score comparing their likelihood of being in a conserved or non-conserved state. It has the specific advantage of not requiring a fixed window size.
Regulatory Potential scores [163,164]	RP scores are derived from two Markov models: the first is trained on known regulatory elements; the second, its neutral model, is trained on ancestral repeats. The model is selected based on optimizing intra-species conservation against

	nucleotide motif signature. It calculates its final score by summing the logarithm of the probability that a state belongs to the regulatory region model versus the ancestral repeat model over a window of fixed length.
DLESS scores [165]	A disadvantage of previous scores was that they biased towards conserved regions that were under selection in all taxa. The DLESS algorithm was designed to identify sequences that may be under selection in a particular lineage. It uses a phylogenetic Markov model adapted from the PhastCons program but introduces the concept of gained or lost sequences on any particular branch of a phylogeny. The final score is a P-value describing the probability of observing an alignment given a neutral model of substitution.

Comparison of MCS, PhastCons, and RP scores, using 93 regulatory regions from the HBB gene complex, have demonstrated sensitivity and specificity in the range of 50-60% [166].

1.3.3.3.2 Transcription factor binding site detection

Detection of transcription factors using phylogenetic footprinting is performed predominantly by limiting transcription factor binding site detection techniques to well-conserved regions. Several computational tools combine database-driven analyses through TRANSFAC or Jasper to detection of conserved elements- among them are ConSite [167], rVISTA [168], Footer [169], CisOrtho [170], and CONREAL [171]. Others, attempt to detect transcription factors *de novo* within conserved regions- among them are Footprinter [172], Monkey [173], and Vestige [174]. Each of these computational tools requires, at minimum, two orthologous sequences which are then aligned using a global alignment algorithm like ClustalW [175] or LAGAN [176]. It has been demonstrated, however, that the efficacy of these algorithms in regulatory regions is dependent on the divergence time of the species being aligned [177-179]. To improve

the ability of alignment algorithms to reconstitute known binding sites, several alignment algorithms have been developed which are anchored based on reconstruction of potential transcription factor binding sites- among them are CONREAL [171], SITEBLAST [180] and TF-map [181]. However, the underlying processes of transcription factor binding site evolution has highlighted lineage-specific gain and loss as a salient force and suggest that the utility of broadly-selected multi-species conservation to transcription factor binding site detection will be limited [182,183]. Examples of such lineage-specific stabilizing selection has been demonstrated in the *eve2* enhancer [184,185] and the Hoxb2a-b3a intergenic region [186].

Recent availability of whole genome data has resulted in further advancement in transcription factor binding site identification through phylogenetic footprinting by facilitating genome-wide surveys of statistically-significant, conserved binding motifs. A specific application in four *Sacchromyces* genomes has been used to identify over-represented motifs in conserved elements [187]. The advantage of this approach is that it uses conserved signals across the genome to identify *de novo* transcription factor binding motifs. By using a genome-wide approach and filtering against functional constraints from expression data, these authors were able to identify 79 potential binding motifs. Section 1.3.3.4 will focus on how functional information like expression data can be used to detect transcription factor binding sites.

1.3.3.3.3 Phylogenetic shadowing

A corollary of phylogenetic footprinting has been the development of the phylogenetic shadowing technique. With the increasing availability of many closely-

related genomes, such as primates, *Drosophila*, *Caenorhabditis,* and rodents, phylogenetic shadowing is a technique which measures conservation by identifying regions of invariance shared among a group of closely-related species. The advantage of this technique is that lineage-specific elements are more likely to be identified and alignments are more informative due to decreasing divergence. Application of this technique in primates has been used to identify 10 protein binding locations in the apo(a) promoter of humans which when compared to lowly-conserved regions which demonstrated increased significance in driving reporter gene activity [188]. To increase the application of this method to novel genomic regions, the authors of this study have made available a tool for phylogenetic shadowing called eShadow [189]. Despite the promises of such an approach, recent insight into population size and corresponding selection pressure has provided evidence that most hominid mutation is neutral, potentially confounding the ubiquity of this approach [190].

1.3.3.4 Gene function-driven approaches (Expression, ontologies and localization)

The major premise of gene function-driven approaches is that genes demonstrating similar properties potentially share similar regulatory elements. These approaches utilize diverse sources of gene function data, most commonly including gene expression, ontological, or localization data. This section will provide an overview of how each type of data has been used to identify regulatory elements.

1.3.3.4.1 Using expression data to identify regulatory elements

The maturation of two technology platforms for measuring gene expression activity *in vivo*, biological chips [191] and serial analysis of gene expression (SAGE) [192], has facilitated statistical determination of sets of genes with common expression patterns. Availability of sets of coexpressed genes facilitate identifying transcription factor binding sites as coexpression implicates that the genes are regulated by a mutual set of transcription factors. The criteria for selecting coexpressed genes are outside of the scope of this thesis and have been reviewed elsewhere (reviewed in [193]). Application of this technique has been widespread, however, with specific usages ranging from identification of regulators involved in environmental response in yeast [194] to cancer in humans [195]. A specific challenge with this approach is that the extent and type of gene expression data included in any regulatory analysis will influence the underlying biological hypothesis for regulatory element detection; by having broad inclusion criteria, stage- or tissue-specific coexpression will likely be masked hiding those regulatory elements that are involved in regulation of a specific biological process. Conversely, too stringent of a gene expression set may mask broader patterns of coexpression and hide regulatory elements coregulated in non-specific processes.

Specialized bioinformatics resources have been developed to aid in identifying regulatory elements using expression data (reviewed in [196]). The oPOSSUM tool takes as an input a list of genes that have been previously identified as coexpressed and then calculates the statistical significance of conserved binding sites within their associated core promoters [197]. Tools like CARRIE, take raw expression data and promoter

sequences to derive a gene coexpression set from which overrepresented binding sites and a gene interaction network are ascertained [198,199].

1.3.3.4.2 Using ontological data to identify regulatory elements

The rapid increase and global distribution of genomics data has necessitated the development of controlled vocabularies, or ontologies, capable of describing shared properties or concepts. An example of these ontologies is the Gene Ontology which established the semantics of describing the different biological roles that genes can possess [200]. An advantage of the Gene Ontology to regulatory analysis is that it becomes easier to extract genes known to be involved in particular metabolic pathways. These genes can then be analyzed for regulatory elements that may be specific to a particular biological role. An example of this type of analysis has been performed to characterize candidate binding sites in immune-response-related genes [201]. More typically, however, the Gene Ontology has been used to validate the results of expression-based regulatory analyses. Two recent studies, however, have utilized database-driven approaches to cluster genes possessing common sets of putative transcription factor binding sites and then have validated these clusters against the Gene Ontology [202,203]. The continued development of new ontologies describing expression data, inclusive of anatomy, cell type, pathology, and development stage [204], and the establishment of diverse ontologies under the Open Biomedical Ontologies foundry [205] will, when sufficiently utilized, further offer enriched vocabularies to aid in identifying regulatory elements essential to particular biological roles.

1.3.3.4.3 Using localization data to identify regulatory elements

Observation of gene locations in several genomes has demonstrated a tendency for co-expressed genes to physically cluster (reviewed in [206]). This is likely because either tissue- or stage-specific chromatin domains allow neighbouring genes to be expressed at the same time and/or they share similar regulatory control mechanisms [207]. The former scenario has been supported by the discovery of chromosomally-clustered genes involved in intestinal and muscle-specific genes in *C. elegans* [208,209]. The latter scenario has been supported by the discovery that 10-20% of the genes in the human genome share bidirectional promoters [210]. The potential for chromosomally-clustered genes to share regulatory elements suggests that future genome-wide approaches cannot simply prescribe to a one gene, one promoter model.

1.3.3.4.4 Other data sources

Any source of information which clusters genes into functional categories may have utility in uncovering regulators involved in driving a particular function. Limited information is currently available describing chromatin state in specific cell types, the regulatory regions driving non-coding RNA transcription or the *in vivo* nuclear structure including transcriptional factories and chromosomal domains. Future characterization of these regions should provide additional resources for investigating regulatory control mechanisms.

1.3.3.5 Architectural features of regulatory regions

Investigations of the general architecture of regulatory elements has uncovered general features of regulatory regions and transcription factor binding sites that can be used as predictors in other regions of the genome. A previously introduced example of such a feature is CpG islands which are correlated with a class of promoters. Other features of regulatory architecture (which will be discussed in this section) are combinatorial or composite binding, where clusters of binding sites are used to elicit regulatory regions; DNA-protein structure, where the elucidation of the structural configuration of the DNA-protein contacts can indicate a regions amenability to binding; and repetitive elements, which can both describe the selection pressure on a region and introduce new regulatory elements.

1.3.3.5.1 Combinatorial (composite) binding of transcription factors

A salient feature of gene regulation is that multiple transcription factors can act together to alter gene expression (reviewed in [38]). This interaction can be cooperative, where an expression change is the sum of the cumulative effects of each transcription factor taken independently [211], synergistic, where an expression change is greater than the cumulative effect of each transcription factor taken independently [212]; or competitive, where an expression change is the output of multiple antagonistic effects (reviewed in [213]). Furthermore, each change in gene expression activity can be the product of many different transcription factors or a multitude of the same transcription factor [214]. Genome-wide analyses of transcription factor binding in *Saccromyces cerevisiae* has elucidated that both such arrangements are widespread [215]. This

architectural feature of regulatory regions has been utilized extensively to predict new regulatory regions. Studies have been designed to detect regulatory regions with a significant density of binding sites for all known transcription factors [216-218], transcription factors which are known to be tissue- or stage-specific [219-224], or individual transcription factors [225-227]. Among the methods described in some of these studies, several tools are available to assist in performing combinatorial analyses, including Cister [223], Comet [228], Cluster-Buster [229], MSCAN [230], CATS [231], and TFBScluster [232]. Most used database-driven approaches to detect regions which contain an overrepresentation of predicted binding sites for a user-selected set of transcription factors. To improve these predictions, databases like TRANSCompel provide information on experimentally validated interactions between transcription factors and their binding sites [233].

An extension of this method has been applied to predicting regulatory elements from their spatial organization or symmetry. Observations that lambda repressor binding sites are separated by integral turns of the DNA helix [234] and MEF2 and MyoD families of transcription factors bind a fixed distances relative to the DNA helical turn [235]. Computational analysis of spatial organization of transcription factors in *Drosophila melanogaster* has elucidated periodic signals for the Bicoid and Hunchback transcription factors [236]. Furthermore, many transcription factor binding sites have been identified as having 'dyad' structures, well-conserved bases separated by several poorly-conserved bases [237]. Many of the previous discussed *de novo* motif discovery techniques are specifically designed to exploit dyad symmetries. A Gibbs sampling technique has been adapted to specifically improve detection of these types of

arrangements [238] by allowing users to investigate different types of motif symmetry. However, while some transcription factors may require specific spatial organizations, it is generally viewed that many do not [239].

1.3.3.5.2 DNA structure and gene regulation

The curvature of DNA is known to have a role in gene regulation in prokaryotes [240,241]. It is largely regarded to be involved in temperature response [242,243] and/or maintenance and positioning of chromatin structure [244-246]. Evidence also suggests that they are responsible for recruiting specific transcription factors [247,248]. Conserved structural motifs have been identified in eukaryotic ribosomal promoters and in satellite and nucleosome positioning DNA independent of sequence homology [249,250] and a scan of eukaryotic promoters has demonstrated a correlation between the GTF, TBP, and DNA curvature [251]. Future characterization of structural motifs may elucidate higher-order regulatory signals or uncover important protein recruitment mechanisms.

1.3.3.5.3 Repetitive elements and gene regulation

Repetitive elements are an architectural feature of genomes that have traditionally been ignored in regulatory analyses because of their predisposition to skewing results towards prediction of the repetitive motif. However, repetitive elements have been demonstrating an increasing relevance to the function and evolution of gene regulation. A diversity of predicted transcription factor binding sites from developmental transcription factors on Alu repeat elements near biosynthesis genes implicates them as

being important to suppression of proliferation during differentiation [252]. Long terminal repeats (LTRs) are also known to contain strong promoters which can drive the expression of individual (reviewed in [253]) or multiple genes [254]. Furthermore, the mutational qualities of such repetitive elements (specifically their ability to add or subtract repeat units) make them ideal targets for both *cis*- and *trans*-acting regulatory variation [255,256]. The potential of repeats to subvert a gene's regulatory program has also made them an ideal tool for investigating the selection pressures in non-coding sequences [257,258]. A better understanding of the role of repetitive elements in gene regulation will likely improve computational identification and characterization of regulatory elements.

1.3.3.6 Aggregate approaches (Workbenches and pipelines for regulatory element analysis)

The ample computational resources at a researcher's disposal for regulatory element analysis have fostered the emergence of aggregate approaches which aim to integrate and present the state-of-the-art. Two types of approaches will be discussed in this section- approaches that provide an integrated framework for analysis and high-throughput pipelines which aim to make available pre-computed predictions of general relevance to a larger community.

1.3.3.6.1 Tools for integrated regulatory analysis

Many of the previously discussed computational techniques for predicting regulatory elements have been integrated into analysis applications combining two or

more different techniques or algorithms. These applications or 'workbenches' have well-defined utility as they allow a researcher to propose, test, and modify hypotheses using enriched data resources and algorithm selections. Workbench toolkits, like Toucan, Theatre, and SeqVISTA, provide environments where researchers can perform phylogenetic footprinting, *de novo* motif discovery, and combinatorial analyses of transcription factor binding sites [259-261]. Applications like TAMO and BEST provide access to a selection of motif discovery algorithms [262,263]. A disadvantage of these applications is that they are typically not extensible to large-scale or high-throughput analyses. The next section will discuss aggregate techniques that are being employed to pre-compute regulatory elements in whole genomes.

1.3.3.6.2 High-throughput regulatory element prediction pipelines

A major goal of gene regulatory analysis is to be able to produce genome-scale maps of gene regulatory architecture (reviewed in [149]). Several approaches have been reported that aggregate multiple data sources and computational techniques and apply them to detection of regulatory elements; representative examples are discussed in Table 4. Many require extensive computational infrastructure and suffer from antagonistic elements of experimental design, including scope of gene expression data included or conservation depths for phylogenetic analyses. With an increasing diversity of techniques and information that must be synthesized, however, these pipeline approaches offer the principal benefit that they are expert-driven and should be able to complement gene-centric assays where the primary researcher is not intimately familiar with the benefits of different computational techniques.

Table 4: Genome-wide regulatory element prediction pipelines for eukaryotes.

Xie et al. [264]	A comparative analysis of promoters and 3' UTRs for human, mouse, rat, and dog. Methods reduced TRANSFAC database to motif clusters (based on sequence similarity) to detect conserved motifs and used a consensus-based mismatch score for detecting new 11-mer motifs.
PAP [265] and oPOSSUM [197]	A comparative analysis of promoters for human and mouse was conducted. Each promoter was scanned using TRANSFAC and Jaspar databases and each transcription factor binding site is weighted due for all potential interactions. Co-expressed genes are inputted to the database to identify transcription factors that likely regulate the set.
PHYLONET [266]	Orthologues for *S. cerevisiae*, *S. mikatae, S. kudriavzevii*, and *S. bayanus* are determined and phylogenetic footprinting is performed using a BLAST-like algorithm [267] to cluster and assemble motif profiles. Input genes are scanned for matching profiles to assemble a regulatory network.
cisRED [268]	Motifs are assessed using multiple tools and scored using a method-independent scoring framework. Orthologues are obtained from EnsEMBL for mammalian species and act as input to the motif discovery pipeline. Expression data is used to validate the motif discovery approach [269].
PReMod [270]	Non-coding human, mouse, and rat alignments are evaluated for their similarity to individual transcription factor binding sites from TRANSFAC. Clusters of putative binding sites within a 2kb interval are identified within these alignments.

1.3.3.7 Performance assessment

A major challenge confronting computational assays for regulatory element prediction is the sparsity of benchmarks for assessing the performance of tools designed to discover transcription factor binding sites. Currently, very few, if any, regulatory regions are comprehensively surveyed for all possible binding interactions, making definitive predictions of sensitivity and specificity speculative and exposing algorithm designers to bias due to overtraining. Assessments of performance have also typically

used generated sets where binding sites have been artificially implanted; however these are likely not reflective of the true natural process. Also, the availability of complementary types of data from sources such as comparative genomics or gene expression assays have meant that regulatory analyses driven by one type of data can be compared against the other. Examples of this type of performance assessment have been reported where coexpression was used to validate transcription factor binding sites identified through phylogenetic footprinting [271,272] and where binding sites detected by coexpression have been assessed against conservation [273,274]. This mutualistic assessment is the basis behind aggregate tools like PhyloCon which use both conservation and coexpression to identify regulatory motifs [275] and many of the aggregate approaches discussed previously. However, for reasons previously discussed, the ability of these approaches to predict transcription factor binding sites is limited by the "genomic scope" of the data; to further illustrate this, muscle-specific predictions of regulatory regions in *Ciona savignyi* (a relatively small eukaryotic genome, assessed by combinatorial analysis and validated with conservation-based approaches) reached a sensitivity of 46.5% and specificity of 88% when compared with known regulatory regions [276].

A recent study of the performance of motif discovery tools has examined the creation of appropriate benchmarks for assessing these types of tools [97]. This study analyzed 13 different motif discovery tools using data sets for fly, human, mouse, and yeast that had known binding sites in their natural promoter, and in artificially generated promoter sequences. No single dataset was definitive in quantifying performance and most tools performed better on yeast than any other organism. They also demonstrated

the significant challenge in creating useful benchmarks including complications with filtering collections of known sites and incomplete knowledge of all known binding interactions in natural promoters. This study formalized several established and new statistics for communicating performance of motif discovery tools (several statistics that are relevant to this thesis are listed in Table 5). A follow-up study has demonstrated that many motif discovery tools are challenged by input sequence size, the heterogeneity of binding site positions, and the overall similarity of the binding sites [277].

Table 5: Performance metrics for assessing regulatory prediction tools.

Sensitivity	The fraction of known binding sites that are predicted.
Specificity	The fraction of binding sites that are predicted in error
Positive Predictive Value	The fraction of predicted binding sites that are known
Correlation Coefficient	Introduced in regulatory analysis by Tompa et al. [97]. This statistic is a measure of correlation between known and predicted sites. A correlation coefficient of -1 indicates perfect anti-correlation whereas a value of +1 indicates perfect correlation.

A performance assessment of phylogenetic footprinting tools has also been recently conducted by the same lab for small motifs in metazoan promoters [278]. Using well-conserved known binding sites, a measure of parsimony was calculated for motif conservation within a phylogenetic tree. Specific challenges with performance assessment included obtaining correct orthologous genes, their associated promoter regions, and aligning them for distant species. Using human, chimp, mouse, and rat alignments, 85% of these conserved regions, containing a known motif, were recovered and, by adding chicken to the alignment, only 27% were recovered using alignment-based tools.

The recent availability of high-throughput biological assays for identifying DNA-protein binding interactions, like ChIP-chip, has the potential to aid enhancer and promoter prediction. Furthermore, recent genome-wide discovery of transcription start sites [40] offers further potential to benchmark and extend the efficacy of promoter prediction tools like Eponine [91] and PromoterInspector [279].

1.3.4 Synopsis: Future trends

There has been marked growth in the types of biological and computational assays available for analyzing gene regulatory properties. International collaboration through the ENCODE project aims to survey functional elements in 1% of the genome; this resource will undoubtedly provide useful training material for existing and future types of assays. Especially since many of the approaches currently used for detecting regulatory elements have been principally designed through training analysis in yeast [97], and some of the more interesting discoveries in lower-order organisms, like coding versus non-coding word bias have remained unexplored in higher order genomes [280]. Fundamental questions still need to be addressed though. There remain relatively few benchmarks for surveying, comparing, and integrating different regulatory methodologies. Very little is known about species-specific contributions to comparative genomics analyses in regulatory regions in light of a plethora of genomes rapidly becoming available. Additionally, very little is known about the contributions of individual expression assays to detecting reliable co-expressed genes. It is certainly

obvious that there are particular caveats when utilizing broad versus condition-specific expression sets which have consequential effects on the types of regulatory elements identified [181]. To address these challenges, the future class of regulatory prediction algorithms will need to be able to more accurately discern selective and functional constraints in a wider biological context. Much remains to be explored, compared, and pragmatically integrated in terms of conservation, coexpression, and the structural features of the genome whether for genome-scale or targeted analyses.

1.4 On bioinformatic analysis of genetic mutation

1.4.1 Background

The diversity endowed in nature is primarily due to DNA sequence variants embedded in each organism's genome. These variants can exist in almost any form: as small base-pair mutations (single-nucleotide polymorphisms), insertions/deletions, and large-scale chromosomal mutations. Each has the potential to alter an organism's fitness and, when located in germline tissue, propagate to its progeny. Cataloguing sequence variation offers the potential to investigate the molecular causes of phenotypic difference and to trace the origins of species [281].

1.4.1.1 Discovering genetic mutation

Genetics is the study of mutation and its inheritance. As such, the roots of our understanding of genetic mutation have had their ascendancy with the history of genetics and the advent of DNA-based technologies. Classical genetics approaches (as those employed by Morgan in identifying the *white* mutation in *Drosophila melanogaster*)

require first the identification of a mutant phenotype, which is then mapped to its associated sequence variant. To aid in obtaining interesting phenotypes, organisms are typically mutagenized and screened *en masse* to select those with identifiable phenotypes. However, challenges with this approach include that it is time-consuming, many mutations do not have a visible phenotype, and many small genes are hard to disrupt. DNA sequencing has heralded the development of reverse genetics-based approaches where, conversely, a sequence variant in DNA is assayed for its corresponding phenotype. The advantage of this approach is that candidate genes, regardless of size, can be selectively targeted and/or specific mutations can be artificially induced to mimic *in vivo* mutations. Reverse genetics has identified genes responsible for rare, highly heritable 'mendelian' diseases, such as cystic fibrosis and Huntington's, by identifying genetic markers (alleles) from DNA sequences that were linked to the respective diseases [282,283]. A historical challenge with this technique, however, is that it has limited utility in identifying the loci for complex or common traits which may be the product of multiple interacting sequence variants. As such, these types of traits have now been studied using genome-wide family-based linkage studies and population-based association studies, which, while effective, are challenged by the ability to significantly detect linkage in multi-locus traits and the extent of variants that can be assayed, respectively. Improvement in genome sequencing technology has and will continue to have a multi-faceted effect on the identification of all types of genetic mutation. Of significance, it has facilitated the extensive identification of common genetic variants, useful for both linkage analysis and association studies, and has promoted the advancement of reverse genetics-based approaches to creation of "knock-out" libraries,

where all the genes in an organism are individually disrupted. Genomics-facilitated discovery and genotyping of many new genetic variants have offered powerful tools for extending our current understanding of the distribution and function of genetic mutation as it pertains to health and the ancestry of a population.

1.4.1.2 Current model of genetic mutation

The most abundant type of genetic mutation in the human genome are single nucleotide polymorphisms (SNPs) (reviewed in [284]). These variants are typically biallelic where one nucleotide has been replaced by another most likely through an error during DNA replication or repair. SNPs are generally catalogued into those which occur in protein coding sequences and those which do not. The SNPs that occur in protein coding sequences are regarded to be more likely to have a role in effecting gene function and are further classified into non-synonymous and synonymous SNPs; the former are widely considered to be more likely to have an effect as they change an associated protein's amino acid sequence in allele-specific manner. Other SNPs likely to affect gene function have been investigated due to their co-localization with splicing sequences [285,286] or canonical regulatory regions [136,287].

The rapid identification of SNPs has been primarily due to the growth of DNA sequencing. The sequencing of the human genome originally introduced approximately 1.5 million single nucleotide polymorphisms distributed at approximately 1 SNP every 2kb; these SNPs were identified by analyzing nucleotide variation in high-quality matching positions of overlapping sequence reads [288]. Within 5 years of this

publication, the quantity of SNPs in public databases has grown to over 12 million SNPs or approximately 1 SNP every 250bp [289].

The large number of available SNPs has made it increasingly feasible to selectively test these variants for their role in etiology of common and/or complex traits. This is because most humans have low genetic diversity compared to other mammals and most of the heterozygosity in the human population has been attributed to common variants (>1%) [290,291]. This observation has stimulated the hypothesis, known as the Common Variant-Common Disease (CV-CD) theorem, that common variants are sufficient for detecting common disease [292-294].

The practicality of using SNPs to detect common and/or complex traits is enhanced by the observation that particular sets of alleles are inherited together more frequently than by chance; this association is called linkage disequilibrium (LD). An allele at one location, therefore, can be informative for the presence of other alleles. The set of alleles that are co-inherited together is commonly defined as a "haplotype". The extent of any haplotype is defined by the extent of LD which is generally influenced by a regions recombination and mutation frequencies, the fitness interactions between genes, and non-adaptive processes, such as limited population structure or inbreeding. Uncovering the distribution of haplotypes has been the focus of an international consortium called HapMap, which has recently genotyped 1 SNP for every 5kb from 4 different populations [295]. By identifying the distribution of haplotypes, it is reasoned that an optimal number of informative SNPs can be selected for association studies. However, as described by Terwilliger et al. [296], there are several caveats with these studies, specifically: 1) the phenotype has a measurable effect, 2) the combined effects of

multiple rare variants are negligible, and 3) the extent of LD in a genomic region from a given population is sufficiently large. Furthermore, it is important to recognize that association does not define a causative variant. While alone this result can be beneficial in its diagnostic utility, until the specific molecular players are identified, further developments of pharmacological interventions are limited. To expedite this, several computational approaches are offering insight into likely causative mutations.

In addition to SNPs, an emerging class of genetic variants, characterized by variable-length deletions, has extended the role of mutation in human populations [297]. Hundreds of large-scale structural variants have been found with median lengths ranging from 500 to 10.5bp and many thousands more microsatellite and minisatellite repeats are distributed across the genome [298,299]. While not discussed any more here, it is likely that these types of variants will have a significant role in defining quantative traits and disease through processes similar to the characteristic repeat expansion found in Huntington's disease.

1.4.1.3 Current strategies for improving our understanding of genetic mutation

While a growing number of SNPs have been identified, a paucity of genotyping information has typically been available. This information is essential for reconstructing haplotypes. However, the availability of this information will rapidly increase. Especially since the HapMap Consortium (which genotyped 1 million SNPs as part of their Phase 1 publication) is currently genotyping nearly 5 million more. The existing and emerging approaches to discovering and genotyping SNPs will be presented in Section 1.4.2. The utility of these assays, though, is dependent on strategies for

computational post-processing this data. Identification, organization, and analysis techniques will be presented in Section 1.4.3.

1.4.2 Biological assays for the identification and characterization of genetic mutations

1.4.2.1 SNP Discovery

The gold standard for the discovery of new sequence variants has been through DNA sequencing. Initial genome-scale identification of SNPs has made use of diverse DNA sources (reviewed in [300]): expressed sequence tags have aided in the identification of nearly 100 000 SNPs [301,302], SNPs were identified from reduced representation shotgun sequencing libraries [303], random shotgun reads aligned to the genome [288], and comparisons of overlapping regions of large-insert clones [304] during the sequencing of the human genome. However, the principal challenge with sequencing-based approaches has been that it has remained relatively expensive.

Alternate SNP discovery technologies have been developed to simplify identification and reduce costs (reviewed in [305]). Two methods that are the most popular are single-strand conformation polymorphism (SSCP) analysis [306] and denaturing high-performance liquid chromatography (DHPLC) [307]. SSCP identifies SNPs based on identifying differences in sequence-dependent conformation patterns of single-stranded DNA under non-denaturing conditions through changes in electrophoretic mobility. SSCP's major advantage is its simplicity, but, among its disadvantages, it requires analyses to be run under different electrophoretic conditions to detect all possible conformational changes. It has sensitivity between 60% and 95% and only DNA segments under 250bp can be assayed. DHPLC identifies SNPs by observing

differences in denaturing rate and column retention of DNA heteroduplexes (complementary DNA molecules containing a polymorphism) compared to homoduplexes. DHLPC has the advantage of being automated and allowing greater sample sizes (from 200-700bp) yet it is slow making it unsuitable for large-scale testing. Also, of further detriment, once a variant has been detected, DNA sequencing is required to locate the variants position. While the above are mentioned because of their popularity, many alternative methods also exist, among them denaturing gradient gel electrophoresis [308], chemical or enzymatic cleavage [309], hybridization to oligonucleotide arrays [310], and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry analyses [311].

1.4.2.2 SNP Genotyping

The identification of genetic variants that are determinants of health requires high-throughput genotyping technologies capable of assaying target populations with sufficient power to identify possible susceptibility alleles. While genotyping is possible with any of the methods used for SNP discovery, the advantage of knowing the location of the polymorphism *a priori* has heralded the development of specific high-throughput technologies (reviewed in [312]). Two popular assays are single base extension assays and 5' nuclease assays. In single base extension assays, a dye-labeled primer is specifically annealed adjacent to a SNP which can then be extended by a single dyed nucleic acid. Single base extension products can be assayed by any fluorescence sequencer. In 5' nuclease assay, or TaqMan assays, differences between hybridization of DNA sequences containing full complementary sequences or those containing a

mismatch at the position of the SNP, are differentially cleaved. The cleaved product releases a 5' reporter dye from 3' quencher dye, and produces a fluorescent signal [313]. A disadvantage of this technology is that specialized probes need to be constructed to optimize the dye reaction. Many alternative methods also exist for high-throughput SNP genotyping. Among them are matrix-assisted laser desorption/ionization time-of-flight mass spectrometry analyses [314], pyrosequencing [315], microarray hybridization [316], and bead-based hybridization assays [317]. The construction of the HapMap was largely facilitated with advances in the latter two technologies. Currently, with microarrays, over 500,000 SNPs can be assayed per experiment [318]. Bead arrays take advantage of fibre-optics and microelectronics technology to construct very dense arrays which are able to generate anywhere from 300,000 to 1.6 million genotypes/day [317].

1.4.3 Computational assays for the identification and characterization of genetic mutations

1.4.3.1 Computational approaches for SNP identification

Computational approaches for SNP identification have primarily focused on mining diverse DNA sources for polymorphisms (reviewed in [300]). These approaches have typically focused on optimizing source-specific specificity and sensitivity issues to maximally identify SNPs from sequencing errors or paralogous sequences. For instance, SNP characterization software like PolyBayes, use base quality values from pooled DNA sequences to evaluate the probability that given sequences are from the same genomic region and also contain a polymorphism [319]. Tools such as PolyPhred and novoSNP similarly focus on identifying polymorphisms from base quality values by identifying

peaks in sequence traces at equivalent heights which are likely due to heterozygosity [320-322]. Furthermore, neighborhood quality scores attempt to classify SNPs based on the respective quality of adjacent sequence [303] and have been integrated in tools like SNPdetector, which employ a rule-based system to model decisions made by experienced SNP inspectors [323]. Fast search algorithms, like ssahaSNP, have also been designed to deal with the quantity of data from sequence read databases while incorporating heuristics for identifying likely polymorphic regions [324] and tools like SNPserver/autoSNP identify SNPs from their redundancy at a site and their amenability to cosegration with other polymorphisms [325]. Each of these methods uses preset thresholds to partition SNPs from sequencing errors or paralogous sequences. A neural network-based approach, which implements dynamic thresholds, has demonstrated marginal performance improvements over tools such as PolyBayes [326]. However, sensitivity and specificity values are particularly hard to compare as performance is associated to SNP frequency and/or the number of overlapping sequences used; depending on these values most can predict, at worst, approximately 80% of the SNPs in any positive control set.

1.4.3.2 Genetic mutation databases

The central repository of polymorphism data is the dbSNP database at the National Centre for Biotechnology Information in Washington, DC [327]. dbSNP curates polymorphism data through submission from researchers world-wide and assembles them into non-redundant entries. Several other databases have appeared with

more specific mandates with respect to genetic mutation, some of these are listed in Table 6.

Table 6: Gene mutation databases.

dbSNP [327]	dbSNP is the largest public resource of polymorphism information. It assembles information from researcher submissions. As of build 126 (May 2006), it contained over 12 million unique SNPs for the human genome and close to 38 million more SNPs for 34 different species.
HapMap [295]	Genotypes, frequencies, haplotypes, and assays for phase I and phase II of the HapMap project. Public Release #19 (Oct 2005) contained over 28 million genotyped SNPs.
HGMD [328]	A collection of genetic mutation that is underlying or associated to human inherited disease. Over 53000 mutations have been curated. This database, however, is not open access.
HGVbase [329]	A curated summary of human DNA variation with a focus on the link between haplotypes and phenotypes. Release 16 holds close to 9 million entries.
ALFRED [330]	A database of allele frequency data from diverse human populations.
OMIM [331]	A catalog of human genes and genetic disorders. Information on specific disease-causing mutation is provided with publication cross-references to PubMED.
SNPeffect [332]	A database of coding non-synonymous SNPs combined with information on the likely functional and physiochemical properties of such mutations. Over 74000 SNPs have been analyzed.
SNPper [333]	Retrieval and SNP analysis database built on top of dbSNP and the UCSC genome browser.
JSNP [334]	Database of common SNPs in the Japanese population. Release 28 (May 2006) contained just under 20000 SNPs with approximately 84000 SNPs with allele frequency data.
dbQSNP [335]	A database of SNPs and associated allele frequencies for polymorphisms in the promoter regions of genes assessed through sequencing and SSCP analysis. Version 13 (August 2005) contained ~9700 SNPs.
CASCAD [336]	Contains candidate SNPs associated with expressed

	sequences for *Rattus norvegicus* and *Danio rerio*. Enriches annotation to be able to discriminate SNPs involved in phenotypic variation of different populations.
topoSNP [337]	Resource for mapping non-synonymous SNPs to 3D protein structures. Integrates nsSNPs from OMIM and dbSNP. Publication release contained 27417 nsSNPs corresponding to 770 protein structures.
rSNP_Guide [338]	Contains information regarding artificial or natural variation effects on gene expression. April 2006 release contained 46 entries.
SeattleSNPs [339]	Project focused on identifying, genotyping, and model associations between SNPs that underlie inflammatory response in humans. 279 genes primarily involved in inflammation, lipid metabolism, and blood pressure regulation have been resequenced and over 31000 SNPs have been found as of June 2006. Genes can be externally nominated for resequencing.
GeneSNPs Database [340]	The NIEHS Environmental Genome Project is a resequencing project for genes involved in disease susceptibility in U.S. populations. As of June 2006, their GeneSNPs database contained close to 83000 SNPs from 593 genes primarily involved in DNA repair, cell cycle regulation, drug metabolism and apoptosis.
Human Structural Variation Database [341]	Database contains large-scale structural variation (LSV), copy number polymorphisms (CNPs), and intermediate-sized structural variation (ISV).
Database of Genomic Variants [342]	A curated catalogue of large-scale variation in the human genome
The Chromosome Anomaly Collection [343]	Database contains cytogenetically visible mutation.

Among those databases not listed in Table 6, several databases have established themselves with a focus on polymorphisms relevant to a specific disease. Among these databases are the Breast Cancer Mutation database [344], Cancer Genome Anatomy Project SNP index [345], and the Cystic Fibrosis Mutation database [346]. The value of these databases is that typically rarer mutations identified from disease populations are

differentiated from common mutations and individual variants can be tested for their role in disease.

To facilitate access to genetic mutation data, genome databases, like EnsEMBL and the UCSC Genome Browser and protein databases, like SwissProt, have integrated this information within their repositories. This has allowed researchers to seamlessly investigate known mutations in genomic sequences and proteins.

1.4.3.3 Computational approaches for SNP characterization

The fecundity of single nucleotide polymorphisms has enticed the development of computational approaches which can characterize variants likely to affect function (reviewed in [347]). The majority of these approaches have been applied to ascertaining the importance of SNPs by assessing their selective constraints through determination of the frequency and age of alleles, and their likelihood of altering a sequence's biological function. Each is discussed below.

ALLELE FREQUENCY

Under the right population demographic history, the frequency of an allele can suggest functional constraint. This observation is particularly apparent in human genes when comparing the allele frequencies of synonymous to non-synonymous mutations as the latter mutations are enriched in low-frequency alleles [348]. However, as alluded to, low- frequency alone does not guarantee functional constraint as a balance must exist between selective pressures, which are not just a function of biological importance, but also of population size and generation time, age of the mutation and mutation rate [349].

To facilitate selection of SNPs with specific allele frequencies, computational tools, like Frequency Finder [350] and SNPper [333], have been developed.

DERIVED ALLELES AND FREQUENCY

An estimate of the age of an allele can be inferred by identifying the ancestral allele in closely-related species and calculating a derived allele frequency (DAF; DAFs are the frequency at which the ancestral allele is observed in the reference species population). A DAF can assess the ability of an allele under positive selection to "hitchhike" to low or high frequency [351]. Cargill et al. were able to demonstrate that a significant proportion of alleles have risen in frequency since the human-chimpanzee divergence to become the major allele in the population [348]. This diagnostic has been used to detect recent positive selection in the immune response genes CAV1 and CAV2 [352]. Furthermore, derived allele frequencies have been used to characterize signals of positive selection more generally across the human genome [353] and within conserved non-coding sequences [354].

Allele characterization based on ancestral allele status has been utilized as part of tools that prioritize nonsynonomous SNPs such as SIFT [355] and PolyPhen [356].

ALTERING BIOLOGICAL FUNCTION

A useful approach for characterizing SNPs has been through predicting the consequences of allele-specific changes to the biological function of a DNA sequence. Nonsynonomous mutations are the most frequently quantified SNPs that are suggestive of allele-specific biological effect. This has been extended in coding sequences to identifying SNPs involved in protein structure changes (such as solvent accessibility,

secondary structure, mass, charge, and hydrophobicity differences, location within beta strands or active sites, and their role in disulphide bridges [357,358]). Several computational resources are available for predicting the effects of SNPs in coding regions using functional information; including SIFT [359], PolyPhen [360], SNPs3D[361], SNPeffect [362], PicSNP [363], JADE [364], topoSNP [337], MutDB [365], PolyMAPr [366], and PupaSNP [367]. Synonymous mutations are a frequently quantified type of SNP for not only identifying protein structural constraints but for their ability to affect gene splicing. Computational tools such as PolyMAPr and ESEfinder attempt to characterize these types of SNPs [368]; tools such as PupaSNP and PolyMAPr characterize a SNPs relative role in regulatory regions by assessing the variation in transcription factor binding site predictions from TRANSFAC.

1.4.4 Synopsis: Future trends

The growth of identification and genotyping of genetic variants, including SNPs, in different populations will provide an extensive resource for understanding the effects of genetic mutation in humans. As projects, like HapMap, continue to define the extent of haplotype structure in human populations, projects, like ENCODE, aim to identify new features of genome architecture involved in biological function. The complementation of these two approaches may significantly impact how SNPs are characterized in the future. Furthermore, variants in regulatory regions are suggested to have a prolific effect on heterogeneity in the human population. Improvements in the specificity of algorithms, which characterize allele-specific effects in non-coding regions, will be useful in selecting candidate causative SNPs from disease-associated variants [369].

1.5 On bioinformatics analysis of regulatory mutation

1.5.1 Background

Sections 1.3 and 1.4 have discussed bioinformatics approaches for detecting regulatory elements and genetic mutations, respectively. This section will highlight the intersection of these approaches as they are applied to identifying functional regulatory mutations. Several reviews have been published on regulatory mutation and their role in evolution and disease [369-373].

1.5.2 Experimental identification and characterization of regulatory SNPs

It has long been postulated that the differences in human and chimpanzee are primarily due to regulatory variation since almost all key structural proteins remain virtually identical between the two species [374]. Furthermore, the recent sequencing of the human genome has identified fewer genes than were originally expected, suggesting a principal evolutionary role of alternative mechanisms, such as gene regulation and alternative splicing [375,376]. The role of these variants in describing the phenotypic diversity within a population has been of much interest. Quantification of the extent of regulatory variation has been predominantly explored using experimental assays designed to detect allele-specific gene expression.

Among the earliest techniques used to detect allele-specific changes in gene expression were reporter gene assays. Specific mutations are introduced to assess their ability to drive expression. This technique has been recently used in three independent cell lines to estimate from a population of 170 genes that 35% contain regulatory polymorphisms [136]. Disadvantages of this technique are that it typically does not have

the power to detect small differences in expression level, is laborious for high-throughput screening, and typically only one cell line or condition is assayed [377]. For verification purposes, many reporter gene analyses independently confirm allele-specific effects by assaying for differential protein binding using electrophoretic mobility shift assays.

Characterization of genes with allele-specific gene expression has also focused on techniques which identify genes with unbalanced expression of allelic transcripts in heterozygous samples [378,379]. By identifying particular transcripts that are consistently expressed more frequently than others, it is assumed that each transcript is in linkage disequilibrium with an associated regulatory polymorphism. This approach has demonstrated its efficacy in identifying lowly-expressed transcripts in monogenic diseases like Marfan syndrome [380]. As part of high-throughput assays in normal tissues and conditions, this approach predicted between 25-50% of genes have allele-specific expression patterns [369].

More recent techniques have taken advantage of advances in the scalability of expression technology and have used microarrays to assess expression levels as a phenotype for linkage or association analyses [381-385]. These studies have detected what are termed expression quantative trait loci (eQTLs) by identifying SNPs in significant linkage or association to changes in expression levels. These types of approaches, though, are not without their caveats as expression technology and cell culture introduce noticeable experimental variation when each study is compared [386].

1.5.3 Computational characterization of regulatory SNPs

Very few computational strategies have been developed to identify regulatory variants despite the advantage of restricted sequence space and the diversity of computational assays available. SNP characterization tools, like PupaSNP [367] and PolyMAPr [366], identify putative regulatory SNPs by comparing their allele-specific predictions from TRANSFAC. This approach has been used for genome-wide identification of polymorphisms disrupting well-characterized consensus sequences; this survey demonstrated significant utility in locating regulatory variants within p53 response elements near the transcription start sites of genes in the p53 response pathway as 8 out of 8 polymorphisms tested demonstrated function [387]. This technique, when coupled with phylogenetic footprinting between mouse and human, has further demonstrated its utility on a set where 7 out of 10 SNPs that showed significant allele-specific differences in Jasper predictions also demonstrated electrophoretic mobility shift differences. However, only 2 of the 7 had marked effect in reporter gene assays [388]. A separate database called PromoLign makes available SNPs and pre-computed conserved binding sites between human and mouse for further analyses [389]. While neither study was statistically definitive, clearly, this suggests limited effectiveness of approaches using database-driven regulatory prediction comparisons alone.

The only alternative strategy published to date was a recent statistical analysis of sequence composition from a collection of known regulatory mutations which recognized that the composition around functional SNPs should be different than that around non-functional SNPs due to their selective roles in transcription factor binding [390]. These authors observed that functional mutations of type C-to-T are slightly more associated

with DNA regions with lower average sequence complexity with respect to symmetric elements and is likely attributed to the known dyad symmetry of some transcription factor binding sites. In promoter regions, this technique was reported to have 70% specificity and 20% sensitivity. However, the ability to discriminate intragenic polymorphisms for functional importance was no better than random which was attributed to the absence of promoter-specific sequence composition differences. Of more consequence, their positive control set was of limited statistical power, as only 44 polymorphisms can be confidently characterized as functional. It is this paucity of known functional regulatory polymorphism and the conjectural identification of non-functional SNPs which significantly challenges our ability to develop robust discrimination techniques.

To address the lack of known regulatory polymorphisms databases like MutDB, EnsEMBL and UCSC show conservation profiles with polymorphism data and databases like rSNP_Guide and HGMD catalogue regulatory mutations. However, the cumulative total of germline regulatory polymorphisms, which cause gene expression changes within the latter two databases, is approximately 60 and it is often difficult to discern what experimental conditions were originally used to predict them. Of note, rSNP_Guide additionally provides software for predicting the effect of a SNP when coupled with user-supplied electrophoretic mobility shift assay data.

Population- or evolutionary-based identification of selection pressure on regulatory polymorphisms offers a complementary approach to these methodologies. It has been demonstrated that allele frequency shifts from human-primate divergence can detect functionally constrained regions [354]. Furthermore, allele frequencies below 6%

when assayed in 114 human genes have been observed to be enriched in functional polymorphisms [391]. Both these studies suggest that frequency-based discrimination of functional regulatory polymorphisms should aid other computational approaches.

A purely computational approach to detecting allele-specific expression difference has been conducted through mining publicly available EST data [392]. SNPs in ESTs that were observed at non-equimolar ratios were assumed to be in linkage disequilibrium with a regulatory polymorphism. When tested, the authors were able to identify allele-specific expression changes in 36% of the genes, not overly different from what would be expected from random gene sampling. But, of significance, the expression changes were common, had been derived from multiple tissue sources, and showed consistency in allele-specific expression results.

A significant challenge to computational approaches aimed at detecting functional regulatory variation is a lack of understanding of the evolutionary history of regulatory regions. The majority of gene expression variation appears to follow a neutral model of evolution which suggests most changes are due to stochastic processes [373,393]. It has been further observed that 32-40% of the human functional sites are not functional in rodents [182]. Specific observations in closely related species of purple sea urchin have identified considerable variation among known transcription factor binding sites compared to flanking sites, suggesting in some situations, sequence conservation is not always necessary for evolutionary retention of function [394,395].

1.5.4 Synopsis: Future trends

Several challenges are of imminent importance to the identification of functional regulatory variants. Standards for describing experimental conditions and controlling experiment variation will be required to facilitate comparison of high-throughput technologies and cell conditions [386]. Advances in computational techniques to prioritize candidate regulatory polymorphisms are required. Specifically, the majority of characterizations strategies use database-driven regulatory analyses approaches which are prone to high false-positive prediction rates. Furthermore, allele-specific analysis of *KRT1* expression has demonstrated that multiple *cis*-regulatory polymorphisms are likely responsible for allele-specific expression differences and future characterization of causative variants will require unraveling the individual contributions of each [396].

1.6 Thesis objectives and chapter summaries

Elucidating the spatial and temporal processes which control gene expression remains one of the principal challenges of biology. It is well understood that a component of regulatory control is governed by how specific transcription factors bind DNA to activate or repress expression of nearby genes. Many computational and biological assays have been designed to identify the sites of transcription factor binding. However, many challenges exist. Among them, very little attention has been paid to the biological context of these methods; specifically, identifying the relative importance of information derived from genome sequences and their diverse annotations. Furthermore, there remains a lack of standards by which computational approaches can be identified and assessed. This challenge is in part due to growth in availability of computational and

experimental assays but also a function of the scarcity of well-defined benchmarks for comparing approaches. To address these challenges, gene regulatory analyses will need to address these challenges and take advantage of richer collections and descriptions of identified binding sites, better genome integration of results as they pertain to functionally-related sequences, and more identifiable benchmarks for utilizing and comparing diverse bioinformatics resources.

The purpose of identifying gene regulatory processes is to understand the effects of dysregulation on human health and evolution. Targeted analysis of non-coding polymorphisms has identified variants associated to cancer [397,398] and genetic conditions, like depression [399], systematic lupus erythematosus [400], perinatal HIV-1 transmission [401], and response to type 1 interferons [402]. Most of these studies, while highlighting the importance of functional non-coding polymorphisms, do not identify the causative polymorphisms. Furthermore, the recent developments of allele-specific expression and whole genome association and linkage approaches have suggested that a considerable fraction of the human genome's heterozygosity is due to regulatory mutation and have further presented strategies for identifying the common or complex determinants of disease. These studies, however, are limited to the resolution of identifying a regulatory haplotype and not the causative polymorphism. The advancement of computational approaches, trained to discriminate functional from non-functional polymorphisms, will aid in prioritizing candidate polymorphisms for experimental testing. A limiting factor of this development has been the scarcity of well-catalogued functional polymorphisms. In addition, despite approaches which undertake phylogenetic-based approaches or sequence composition-based approaches to

discriminating functional variants, no approach has looked at the relative contributions of diverse regulatory properties in combination with population-based signals of selection.

The primary aim of this thesis has been to improve strategies for identifying, comparing, and visualizing regulatory element to be able to identify intrinsic properties of SNPs affecting regulatory elements. This thesis can principally be broken down into two major sections where Chapters 2, 3 and part of 4 discuss research undertaken for the purpose of characterizing regulatory elements and Chapters 4 and 5 discuss the development of a technique for characterizing regulatory polymorphisms. The first section of this thesis describes the development and utility of four bioinformatics resources: Sockeye, Chinook, ORC and ORegAnno. This second section addresses causative regulatory variants as integrated through ORegAnno and an approach for using diverse regulatory properties and population genetics features for discriminating them within a promoter sequence. A summary of the analyses as presented in Chapters 2 through 5 are given here.

In Chapter 2 (as published in [403]), I describe a bioinformatics resource called Sockeye which was designed to integrate and visualize functionally-related genome sequences and their annotations to improve identification and characterization of regulatory elements. Sockeye was designed to permit database-driven, phylogenetic footprinting, coexpression-derived, and motif scanning approaches all in the context of EnsEMBL-curated genome annotation. Sockeye was specifically linked to dbSNP to be able to identify locations where putative regulatory elements collocated with polymorphism data.

In Chapter 3 (as published in [404]), I describe a bioinformatics resource called Chinook and an implementation of such for assessing motif discovery algorithms called the Open Regulatory Competition (ORC). With limited resources to perform high-throughput analysis of putative regulatory-SNPs visualized in Sockeye, I was interested in identifying and increasing the efficacies of utilized regulatory prediction tools. To address this, I created Chinook to run diverse sets of algorithms within the Sockeye tool. Chinook was expanded to use peer-to-peer technology to both: allow researchers to connect their own algorithm to Sockeye so that their results could be visually compared against other imported algorithms and annotations; and so that I could also compare different regulatory analyses in the context of available variation annotation. Chinook was used to develop an online platform for comparing these tools using a web application ORC. ORC compares discovered motif discovery algorithms using criteria published by Tompa et al. [97].

In Chapter 4 (as published in [405]), I describe a bioinformatics database called ORegAnno which was designed to aid in curating known regulatory elements and their functional variants. This database was designed to be an open repository for this information and has since led to an internationally-funded, multi-centred collaboration called RegCreative aimed at further populating this resource. Of specific importance to this thesis, within ORegAnno, I hand-curated from 97 publications over 160 regulatory polymorphisms that were identified to individually cause changes in gene expression. This required manually filtering an extensive number of publications that identify regulatory polymorphisms which are only associated to gene expression changes (not confirmed as causal).

In Chapter 5 (in preparation), I describe a bioinformatics approach and associated software called the Cis-acting Human Mutation (CHuM) modules designed to use multiple regulatory and population genetic features of known regulatory polymorphisms from the ORegAnno set to prioritize candidate regulatory polymorphisms for testing. I characterize the importance of particular discriminatory features. I also discuss this method's relevance to an ongoing cancer study at the CMSGSC.

Each of the bioinformatics-based resources and approaches presented in these chapters addresses a broad range of experimentation. The applicability of Sockeye, Chinook and ORegAnno to other investigations will be highlighted in their respective chapters. The extension of the ORC approach to other bioinformatics domains is a tangible alternative. The aim of this work, however, has been to contribute to our overall understanding of gene regulation and the role of genetic variation while critically evaluating alternative approaches to how bioinformatics resources are identified, compared and visualized among peers. Each of these chapters, while unified in their direction, will describe independent contributions designed to achieve this aim.

During the course of my thesis, I have had the opportunity to be involved in several collaborative projects at Canada's Michael Smith Genome Sciences Centre (CMSGSC) which have been described in publications or submitted manuscripts. I have been able to utilize the Sockeye tool to aid in investigation of the SARs coronavirus genome as it was being sequenced at the CMSGSC. At the time, there was speculation as to whether the SARs virus was the recombination product of two other coronavirus genomes; I was able to help provide evidence that revealed that it was not derived from such an event [403,406]. Furthermore, I have been involved in work to organize and

collate scientific publications within the CMSGSC by aiding Martin Kryzinski in cross-referencing publications against PubMed [407]. Finally, I have been involved in a genome-scale regulatory element prediction pipeline called cisRED through initial design and critical evaluation [268].

Chapter 2: Sockeye: A 3D Environment for Comparative Genomics

A version of this chapter has been published:

Montgomery, S.B., Astakhova, T., Bilenky M., Birney, E., Fu, T., Hassel, M., Melsopp, C. Rak, M. Robertson, A.G., Sleumer, M.C., Siddiqui, A.S., and Jones, S.J.M. 2004. Sockeye: A 3D Environment for Comparative Genomics. *Genome Research*. **14(5):** 956-6

Coauthorship details: Dr. Steve Jones and I were responsible for the initial design of the Sockeye resource. Pre-publication, I was a major facilitator and implementer of the Sockeye project. Of note, I designed and implemented several user-interface components and components to embed different bioinformatics analyses, obtain EnsEMBL annotation (such as genes, repeats and SNPs), and visualize regulatory analysis scores. I wrote the published manuscript.

2.1 Introduction

When dealing with biological data, the method by which the data is presented affects the inferences that a trained observer can make. This is especially true when dealing with sequence and annotation data. Users are presented with a wide variety of predicted and experimentally supported annotations from which hypotheses on the structure and function of the sequence can be inferred. Common inferences include whether or not the sequence contains a gene and, if so, what are the sequence features that predict likely function. To analyze these, researchers have had available many genome browsers, some of which target specific organisms [408,409] or groups of organisms, like EnsEMBL [21] and the UCSC Genome Browser [20]. These browsers are extremely well-maintained and offer extensive annotation. However, each browser is designed to reflect information in the context of a limited number of sequences. To address questions such as whether a sequence shares similarities with several other sequences, most genome browsers would require the user to open several independent browser windows. This complicates the perusal of annotation information. Motivated by a desire to perform comparative genomics analyses in an integrated environment, we have designed a software application named Sockeye that allows a user to simultaneously visualize and manipulate both small and large sets of sequences and their annotations.

Sockeye's integrated environment removes the organismal boundaries common in genomics browsers at a time when comparative genomics analyses are positioned to answer questions related to gene regulation and the evolution of structure and function in the genome [152]. Users of Sockeye are able to analyze large sets of functionally-linked sequences, sequences containing genes that are coexpressed and sequences that are

orthologous across multiple species. These sequences can be imported with user-defined annotation or EnsEMBL-curated information, allowing researchers to compare conserved structures. Loaded sequences can be exported, aligned, or discarded. Furthermore, by allowing a user to import custom annotation, Sockeye can facilitate comparative analyses across sequences from any source. For example, we have used this functionality of Sockeye to compare the genome sequence of the SARS virus against the protein complement of other similar coronaviruses [406].

The nature of genome browsing in Sockeye is different from existing browsers. Instead of representing data as 2D images, where annotation is marked above and below the sequence, we have developed a 3D environment where annotation can exploit a 3rd dimension. A 3D environment is conducive to comparative analyses where sequence, organism classification, and annotation take up individual dimensions. In particular, a 3rd dimension (z-axis) is well suited to displaying the magnitude of an associated score when a sequence annotation is predicted *in silico*. Each annotation in Sockeye is displayed as an individual 3D model mapped to a user-defined size and colouring scheme. Each 3D model is specified in a user-configurable XML format file. This allows a user to specify an extensive number of individual annotation objects from only a small collection of 3D primitives (spheres, cylinders, cones, etc.). We propose that the large combination of differential coloring and modeling schemes within a 3D environment will allow researchers to quickly visualize and identify potentially important conservation patterns in multiple sequence datasets.

Sockeye allows users to execute and visualize sequence alignments. Alignment visualization tools like PipMaker [410], while capable of rapidly creating high-quality

images, require set-up of annotation and sequence files. Sockeye integrates the process of obtaining sequence and annotation data. Furthermore, alignment visualization tools like PipMaker, VISTA [411] and synplot [412] offer limited links between alignment results and annotation data (VISTA allows users to view the UCSC site in their Internet browser). Sockeye provides its alignment results with rich annotations like low complexity repeats and ESTs. Sockeye also allows a user to simultaneously visualize several different alignments and easily view their underlying gaps. This allows a user to compare results from an individual algorithm with different input parameters, or from several different alignment algorithms (see Figure 3). Sockeye's focus on providing users with the ability to run genomic analyses, like alignments, defines the principal difference between it and the Apollo project [413]. Apollo, is designed to give a user extensive information to analyze previously run alignments in the pursuit of identifying genomic annotations, like genes. Sockeye, allows a user to perform alignments as required to examine the relationships between sets of sequences when searching for functional non-coding elements.

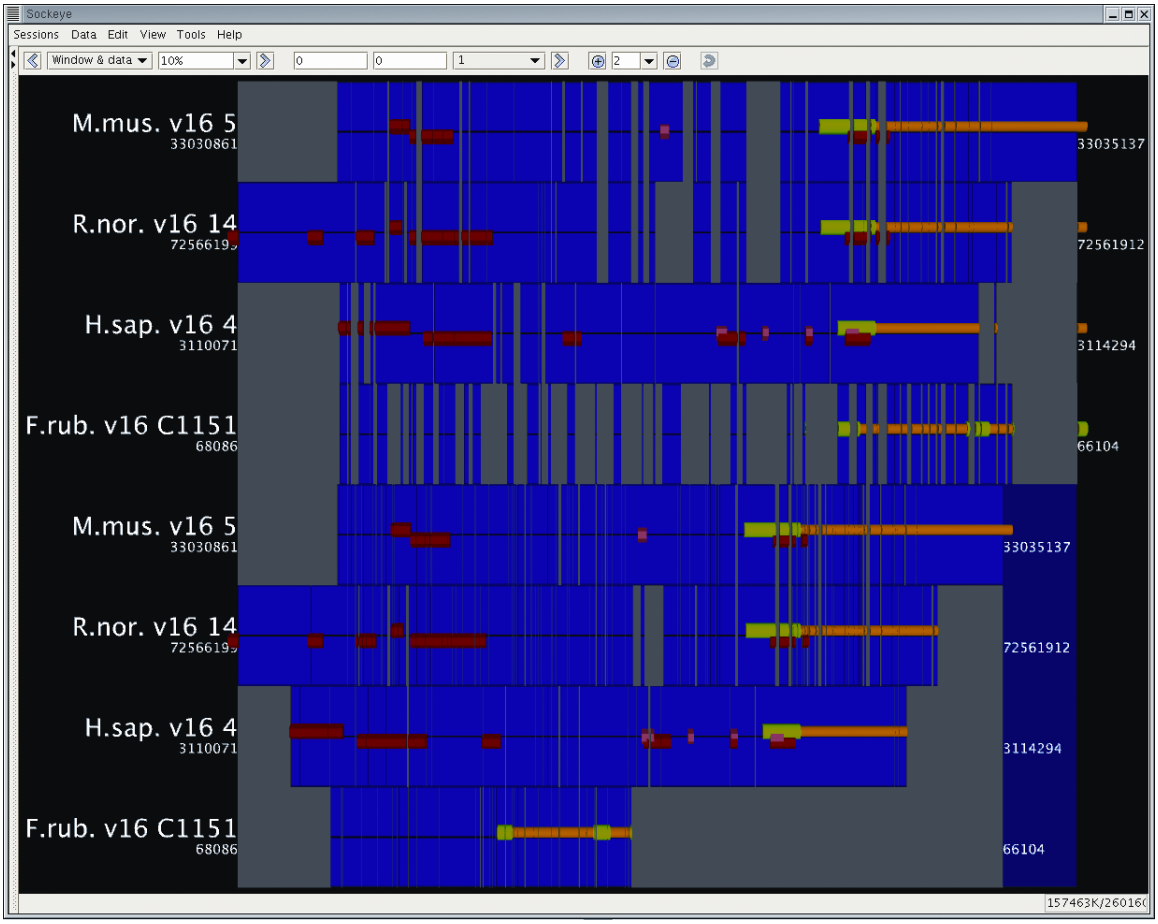

Figure 3: Multiple alignment visualization in Sockeye.

Sockeye is simultaneously showing alignments for regions around the first exon of the Huntington's Disease Protein (HD) from LAGAN (top four sequence tracks) and ClustalW (bottom four sequence tracks). The sequence tracks in order are 1) Mouse, 2) Rat, 3) Human, 4) Fugu, 5) Mouse, 6) Rat, 7) Human, 8) Fugu. For the LAGAN algorithm, the start of translation of the HD gene in each species aligns perfectly. For the ClustalW algorithm, the start of translation for the first exon of the Fugu orthologue is misaligned. The EnsEMBL exon annotation includes the untranslated region. EnsEMBL-curated repeats are shown in red; gaps are shown in gray. This image shows how Sockeye can be used for comparative genomics and comparative algorithmics.

The use of 3D in any type of data visualization and analysis application has advantages and disadvantages. 3D data visualization has been successfully utilized for engineering design and for geological and meteorological modeling but it is relatively unexplored in genomics. The success of 3D applications in genomics has been restricted to those used for protein structure analysis [414-416] and gene expression mapping [417]. Several new applications have emerged to bring 3D visualization to genomic data

analysis. The most notable are the University of Calgary's Java3D CAVE application [418] and the complementary Stichting Academisch Rekencentrum Amsterdam 's Saragene CAVE [419]; both of which use specialized hardware environments, the cost of which is prohibitive for the majority of researchers. Additionally, a more significant disadvantage with 3D applications in general has been that most researchers find it difficult to make quantitative assessments of 3D data as the environment can either enhance or detract from certain characteristics based on an individual's choice of projection. We have aimed to mitigate this problem by providing the user with access to the full range of controls available in most genomic browsers for navigating along sequences and zooming in and out.

One of the fundamental challenges of developing genomic tools and databases is providing the computational interface for research biologists. To approach this, we have focused on providing analysis algorithms to experimentalists without requiring a steep learning curve in applying the algorithm or onerous prerequisite installations. Sockeye allows a user to navigate homologies and align sequences in an integrated environment. Qualitative assessments of sequence similarities can rapidly be made and sequences can be viewed, copied, or exported. Our goal has been to provide the user with an easy system for locating and extracting interesting targets from comparative genomics analyses for subsequent lab and computational study.

2.2 Methods

Sockeye is a standalone application written in Java using JDK1.4.x and Java3D 1.3.x. The EnsEMBL-Java API is used for EnsEMBL-related data access and connection

management. Biojava is used to support file importing and the handling of annotation data. Java RMI is used to connect to standalone analysis programs on the CMSGSC servers. Currently, we use both the JOX (http://www.wutka.com/jox.html, JOX) and XOM (http://www.cafeconleche.org/XOM/, XOM) toolkits to parse our XML startup files and saved files, repectively. Team development occurs in both Linux and Windows via JBuilder and Eclipse. Version control is accomplished using CVS. Sockeye has been tested in OpenGL and DirectX modes with GForce2 MX graphics accelerators. Suggested start-up RAM is 256MB-512MB for large data queries. We have been able to run Sockeye using 64MB of RAM for smaller data queries (~1Mb).

To facilitate development we manage an EnsEMBL data mirror at db01.bcgsc.bc.ca and web mirror at ensemb01.bcgsc.bc.ca:8082. The EnsEMBL data is stored in a MySQL database served from an IBM X440 server with 8-1.5Ghz Xeon processors and 8 Gb of RAM. For presentation purposes, we have been able to mirror EnsEMBL and run Sockeye on a 1.8 Ghz P4 IBM ThinkPad with 60 Gb of hard-disk space and 128 Mb of RAM. The easiest method of using Sockeye is to import data from EnsEMBL at kaka.sanger.ac.uk or db01.bcgsc.bc.ca.

We build our auto-installers for Linux and Windows operating systems using InstallAnywhere. Installations of Sockeye require at least 80Mb of available hard drive for Linux and 55Mb for Windows. It is also recommended that individual users have at least 64MB of available RAM. Sockeye comes prepackaged with the latest Java Runtime Environment and version of Java3D.

2.3 Results

2.3.1 Sockeye Design

The central theme of Sockeye development has been to construct a generalizable and portable software application capable of analyzing and comparing the characteristics of several EnsEMBL-based genome annotations simultaneously.

To facilitate comparative genome analysis, Sockeye is designed to connect and retrieve all major eukaryotic genomes for which a publically-available EnsEMBL annotation exists. A user is currently served annotations of *H. sapiens, M. musculus, R. norvegicus, C. elegans, C. briggsae, D. rerio, F. rubripes, D. melanogaster* and *A. gambiae* through connections to public EnsEMBL MySQL servers. Sockeye's 3D environment is designed to allow the user to simultaneously browse and visualize these annotations. This allows the user to easily view detailed comparisons of genomic structure across multiple organisms. Additionally, to provide genomic annotation in highest possible context, the user is able to connect to relevant external websites for gene annotations. Sockeye currently connects to WormBase [18], Human MapViewer [420], and LocusLink [421].

The design of Sockeye facilitates combining comparative genomics and comparative algorithmics. Sockeye connects to the Chinook application server using Java RMI technology [404]. This service allows a user to run a suite of alignment programs without requiring local installation. The publication version of Sockeye supported ClustalW [175] and LAGAN [176] alignments. The current version of Sockeye has an enriched set of alignment, primer prediction and motif discovery tools provided to it through Chinook (this technology is discussed more in Chapter 3).

2.3.2 Data Structure - Storing and managing retrieved annotations

The fundamental challenges of storing and managing the visualization of genomic annotations are primarily attributed to the complexity of the information and large volume of available data. To meet these challenges, Sockeye creates simple annotation objects called TrackFeatures. These objects are based on Biojava's current version of the GFF specification [422]. In Sockeye, they represent everything from genes to SNPs. The value of this approach is that Sockeye uses only a minimal amount of memory for an annotation. Because of this design, for example, Sockeye allows the simultaneous visualization of all the genes on the six *C. elegans* chromosomes on a P4 workstation with 256Mb of RAM.

2.3.3 Sockeye - A GUI Perspective

On initial start-up, the user sees the sequence track selection tree, the feature selection tree, several navigation controls and the 3D viewport with one empty sequence track visible (see Figure 4). Before any of these components become useful, sequence tracks must be loaded, queried, or imported. A user can use the sequence track selection tree in the upper-left corner to select sequence track specific operations; for example, reversing the strandedness of the visible sequence track, toggling the visibility of the sequence track, or lining-up marked regions. Each sequence track in the sequence track selection tree contains detailed information about its sources; this information can be accessed through the right-mouse button. The feature selection tree allows the user to select which genomic annotations are visible. For example, a user who selects "multitranscript" from the "Gene and gene prediction" branch will be able to see a

floating sphere above the first exon of a gene that contains more than one spliceoform. From these 3D objects, the user can subsequently view the splicing structure of these transcripts or view the distribution of these types of genes across a sequence contig or chromosome.

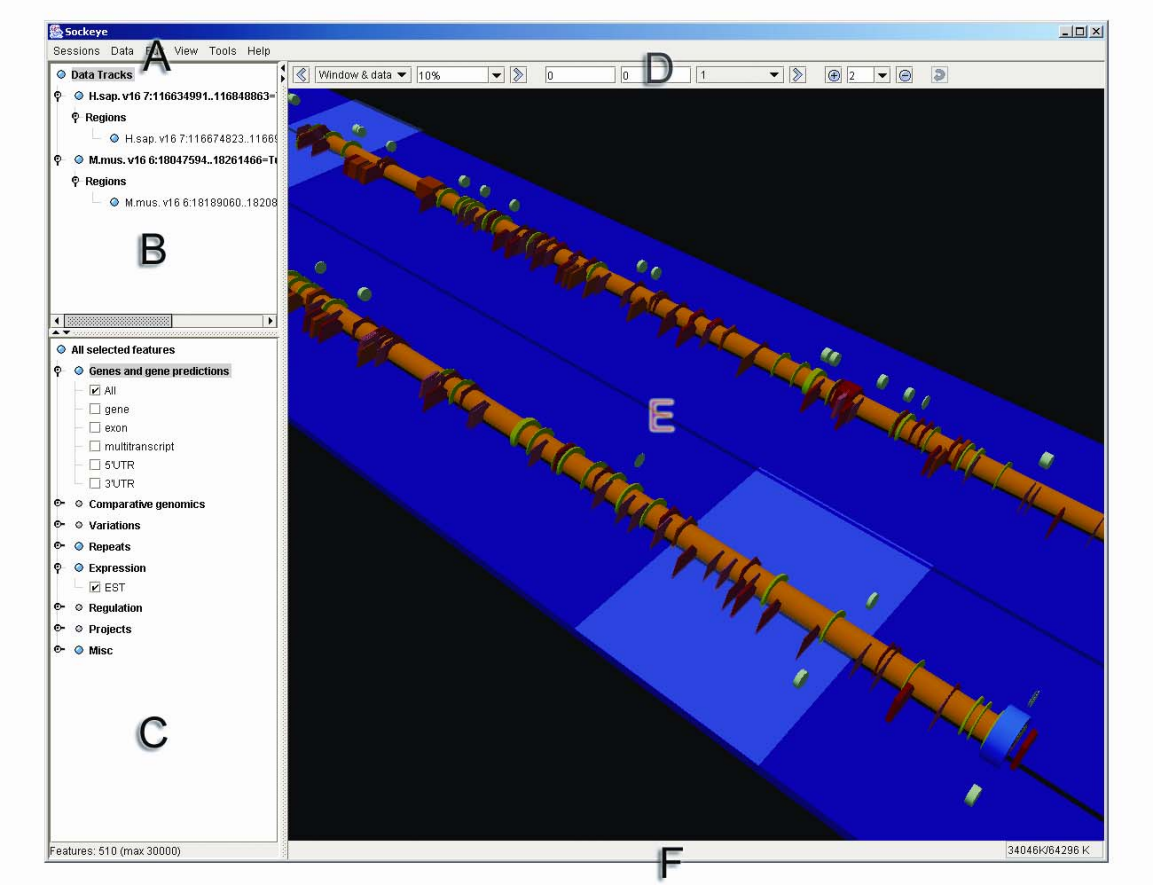

Figure 4: Sockeye GUI layout.

A) The menu. B) The sequence track selection tree. From this component, a user can show/hide and obtain detailed information for loaded sequence tracks. C) The feature selection tree. This component allows a user to show/hide annotation types. It's hierarchical structure is dynamically generated from Sockeye's XML start-up files. D) The navigation toolbar. This component contains tools to navigate loaded sequence tracks. E) The 3D viewport. This is where a user is able to perform analysis and visualization functions on 3D sequence tracks and annotations. F) The status bar. This informs the user of the status of pending EnsEMBL queries and of memory usage.

2.3.4 3D Viewport

Sockeye has been designed to allow researchers to easily compare the extensive information contained across multiple genomic sequences. Sockeye's 3D viewport is where a user is able to visualize and manipulate sequences and annotation data in 3D. Individual sequences are displayed in the viewport as a sequence track (blue plane). As a user subsequently imports additional sequences, the viewport shows these by drawing new sequence tracks adjacent to previous ones. Each sequence track is delineated by its name and start and ending coordinates in nucleotides.

The 3D viewport has been designed to give a user the flexibility in arranging and visualizing data. Once a sequence track is imported in Sockeye, either through an EnsEMBL database query or by loading a file, the user can zoom, pan, and rotate the position of the sequence track using their mouse controls. The sequence track is also an active object; a user who selects the sequence track can highlight or mark individual regions for further analysis. This allows the user to easily select regions of a sequence for refined analyses such as being used for sequence alignment, for sequence export, or for copying into a new sequence track. Additionally, a user is able to adjust the position of individual sequences with respect to annotations on other sequence tracks. This allows a user to import multiple sequences but adjust them to the same reference, such as, the start of a particular gene. This feature has allowed us to line-up homologous genes, aiding the visual identification of upstream similarities.

Individual users can modify some of the display properties of the 3D viewport through the options dialog in the Edit menu and through menu items in the View menu. The dialog allows the user to toggle anti-aliasing and mouseover pop-ups. The View

menu allows the user to change the visible sequence track orientation between a top view and various preselected viewing angles. A user is also given the option of saving a particular 3D viewport orientation so that they can return to it at a later time. This can be particularly helpful if several orientations display important information.

2.3.5 Integrated Support and Maintenance

A major challenge in Sockeye development is that the application had to be easy to use and it would have to be able to adapt to the rapidly changing needs of its user community. Sockeye handles the first of these issues by containing its own context-sensitive help documentation. As well, Sockeye provides its own error tracking and feature suggestion mechanism. A user can fill in reports of erroneous behaviour or desired improvements directly in the application. These reports are sent to a Sockeye issue tracker.

Sockeye users can subscribe to development (sockeye@bcgsc.bc.ca) and announcement (sockeye-announce@bcgsc.bc.ca) mailing lists from the CMSGSC's Sockeye page (http://www.bcgsc.ca/gc/bomge/sockeye, Sockeye). This web-page also includes detailed "How-to" documentation and an online version of the Sockeye help pages.

2.4 Discussion

Sockeye has been designed to leverage 3D graphics technology with tools for performing comparative genomics. We have hypothesized that such an information-rich 3D environment will allow us to quickly view the underlying characteristics of multiple

sequences within a single genome or from multiple genomes. We believe that the success of this design can be demonstrated from ongoing and previous applications of Sockeye to gene regulation, gene discovery and viral genomics studies.

1) The NISC Comparative Vertebrate Sequencing Data: We have used Sockeye to visualize the upstream region of the *CFTR* locus. BLASTn [267] was used to anchor each individual species' contigs to the human genome prior. Subsequently, we used LAGAN to complete a multiple alignment of all the sequences that possessed one BLASTn match to the human CFTR locus. We were able to observe regions of similarity across 14 species.

2) Transcription Factor Binding Site Analysis: We have used Sockeye to visualize the locations of muscle and liver specific regulatory modules upstream of CACNL1AS in human and its ortholog, *Cacna1s*, in mouse. We used a LRA algorithm to scan sliding windows of 200bp incrementing by 5 bp to generate GFF data that was imported into Sockeye as a distribution [220]. From this data we were able to visualize the location of several tissue specific regulatory modules (see Figure 5).

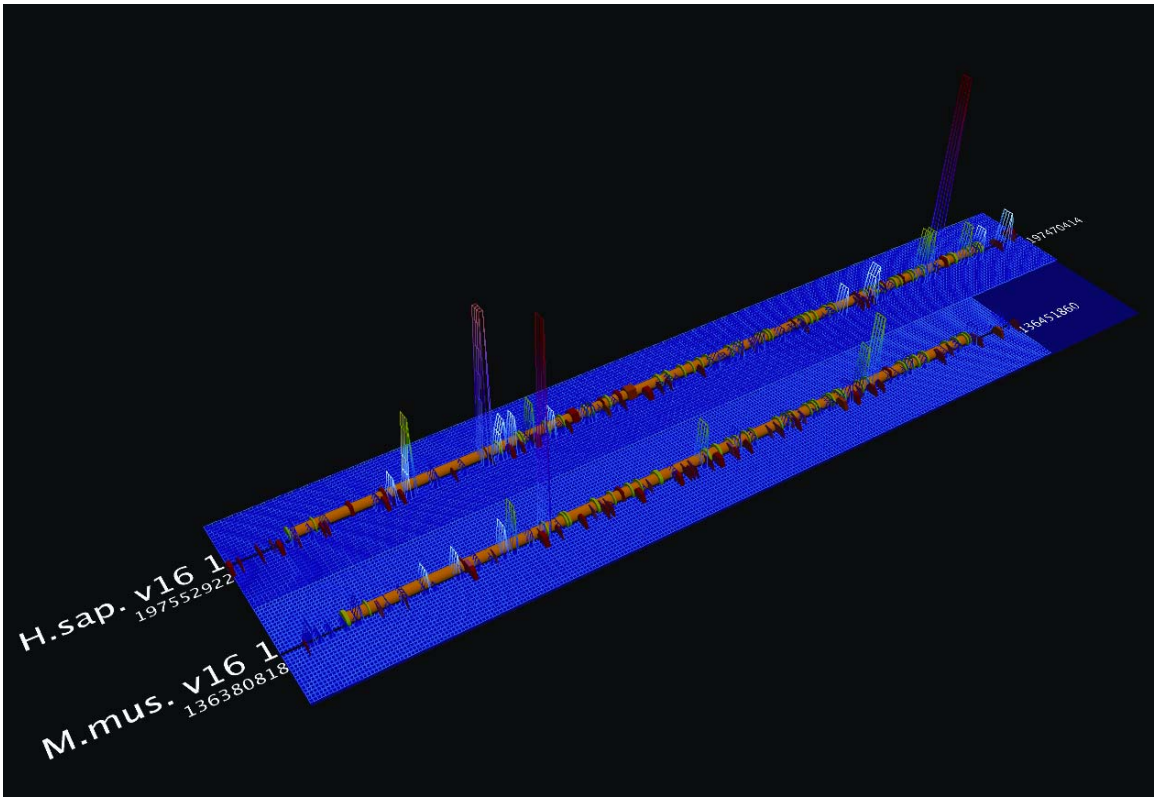

Figure 5: Comparison of Muscle Specific Regulatory Modules in CACNL1AS/Cacna1s in Human and Mouse.

Sockeye displays LRA predictions for muscle regulatory modules as a 3D grid, superimposed on genes (green exons, orange introns). Height and colour intensity represent confidence that a specific region binds muscle-specific transcription factors. To obtain this image, we imported LRA score predictions as GFF and then created a distribution feature in Sockeye's import dialog.

3) Searching for new genes in *C. elegans*: Sockeye is being used in preliminary analyses at the CMSGSC for mapping SAGE tags to *C. elegans*. Because Sockeye shows EST information and *C.briggsae* annotation information, users have been able to find SAGE tags that map to unannotated regions of *C.elegans* and that also share strong homology with annotated genes in *C. briggsae* (G. Vatcher, personal communication).

4) SARS-CoV phylogeny: Using data generated from a BLASTX analysis of the protein complement of the *nidovirales* family against the SARS-CoV virus, Sockeye was able to simultaneously display the dissimilarities present between SARS-CoV and viruses

in this family (see Figure 6). This analysis added further visual evidence that SARS-CoV didn't originate from a recombination of two coronaviruses and was in fact a new type of virus [423].

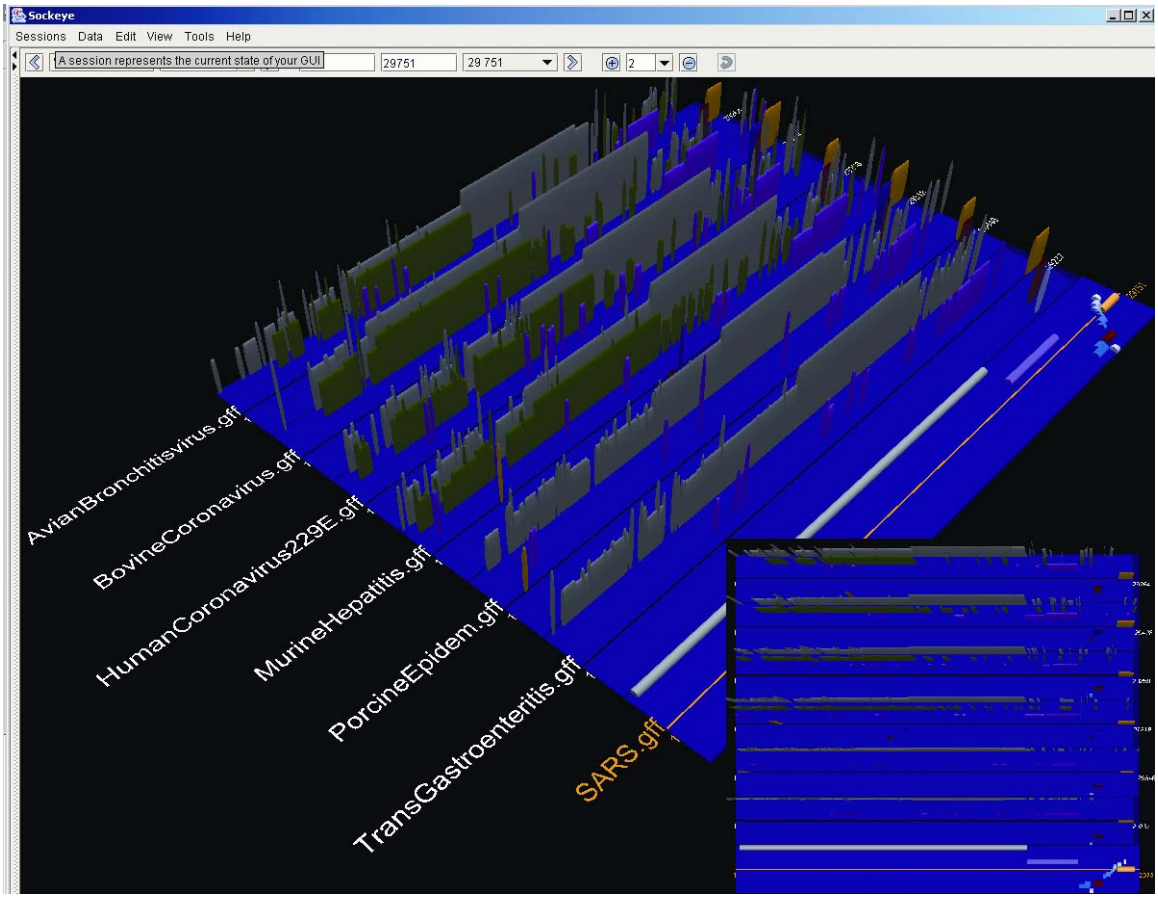

Figure 6: SARs-CoV analysis.

These sequence tracks show a BLASTX analysis of the SARS-CoV virus against 1) Avian Bronchitis Virus, 2) Bovine Coronavirus, 3) Human Coronavirus 229E, 4) Murine Hepatitis, 5) Porcine Epidemic Diarrhea Virus, 6) Transmissible Gastroenteritis Virus. The last sequence track is SARS-CoV annotation imported from NCBI. This image shows dissimilarities between the SARS-CoV virus and other related viruses. These dissimilarities are pronounced in the structural proteins at the 3' end (right). The score for each BLASTX hit is represented by height for each corresponding 3D feature in Sockeye. INSET: The top view clearly shows the nature of the alignment.

The underlying method for completing these analyses was to convert data into a format that Sockeye could read and then import it into Sockeye for visualization. In each

of the above cases, we used GFF files to import new annotation. New 3D models were specified for these features in our XML configuration files allowing us to choose colors and shapes that enhanced the visibility of the data.

Sockeye is well-positioned to handle new types of comparative genomic analysis for gene regulation studies. By allowing a user to set-up and run alignment programs from within Sockeye, we are able to perform phylogenetic footprinting analyses while keeping all EnsEMBL annotation in the context of the alignment. This utility is in contrast to most alignment applications which require the user to run alignments and then determine if similarities overlap known annotations. By integrating this feature into Sockeye, we can easily answer whether known polymorphisms overlap putative regulatory regions and could possibly contribute to an observable phenotype. Additionally, Sockeye allows the user to highlight alternative spliceoforms. This utility is particularly useful for genes with similar regulation mechanism as those found in *Nspl1*, where a specific spliceoform is likely regulated by an intronic promoter [424]. Furthermore, any type of prediction algorithm that can report sliding window scores or putative regulatory regions can be imported directly into Sockeye for visualization as individual features or as distribution data. Furthermore, recent integration of Sockeye with the coexpression set of Griffith et al., allows for comparative analysis of the gene sets [269]. This allows Sockeye to be a workbench for viewing data across functionally-linked genes through orthology or coexpression.

Sockeye is distributed solely as a standalone application. The benefit of a standalone application is that Sockeye, without requiring the mirroring of external databases, can save data while connected to the Internet for perusal at a later time.

Additionally, by supplying XML configuration files in a standalone distribution, Sockeye allows a user to easily customize their own installation. In contrast, however, most genome browsers have been successful because they can easily deliver their content through the Internet to a user. Since Sockeye is a standalone application, it is limited by the CPU, memory and graphics card that a user possesses. A recent version of Sockeye attempts to alleviate these restrictions by incorporating semantic zooming. This mechanism shows information in a form appropriate for the resolution of any given track; for instance, at a certain count and resolution threshold, Sockeye will display SNPs as distributions across the track instead of discrete features. Furthermore, the querying of external MySQL servers can be slow during peak usage; some queries, depending on size, take several minutes. To alleviate this, we have switched from a "query all annotations" strategy to a strategy where we query individual types of annotation on demand.

The Sockeye software is available for Windows (DirectX and OpenGL modes) and Linux (OpenGL mode) download at http://www.bcgsc.ca/gc/bomge/sockeye/. The Sockeye source code is available from the authors under license, at no cost, for academic use.

2.5 Conclusions

While amenable to multiple uses, my focus in this work was specifically to develop an integrated strategy for investigating regulatory elements and regulatory polymorphisms. Sockeye integrates comprehensive annotation, comparative analysis and motif discovery tools along with information regarding alternative transcripts,

coexpression and orthology for any particular gene. This approach is certainly tractable for researchers investigating polymorphisms in a particular gene but, like most genome browsers, is not particularly well-suited to genome-wide analyses. Furthermore, while developing Sockeye, it became apparent that the diversity of bioinformatics tools available for alignment and motif discovery was unwieldy. To address this, I began exploring novel ways of integrating and assessing these tools; this research is the focus of the next chapter.

Chapter 3: An application of peer-to-peer technology to the discovery, use and assessment of bioinformatics programs

A version of this chapter has been published:

Montgomery, S.B., Fu, T., Guan, J., Lin, K. and Jones, S.J.M. 2005. An application of peer-to-peer technology to the discovery, use and assessment of bioinformatics programs. *Nature Methods*. **2(8):** 563

Coauthorship details: Several of the coauthors added small pieces of code to the Chinook project. Of note, Tony Fu provided a prototype of a Web Services implementation. I was primarily responsible for the design and implementation of this resource and the Open Regulatory Competition website. I wrote the published manuscript.

3.1 Introduction

Bioinformatics techniques are used to manage large amounts of biological data and elucidate complex biological relationships. The assembly of multiple genomes and the development of high-throughput assays has made necessary large sources of computational power to analyze and store these relationships. This need is particularly relevant in research domains such as protein structure prediction [425-427], expression profiling [428-430], gene regulation analysis [73,76], genetic analysis [431], and evolutionary analysis [160]. Because of the applicability of bioinformatics techniques to these domains, the ability to rapidly use and compare state-of-the-art bioinformatics algorithms has well-defined utility. However, in many situations, it is difficult to use these algorithms; even when a suitable algorithm has been determined, many run only on specific operating systems, have complex installation requirements, require high-end CPU or other hardware resources, or require the user to adjust to disparate input requirements and modes of operation. To help researchers find specific types of algorithms, many websites [432] and publications [75] have collated links to several top bioinformatics resources; but new integrative efforts have been required to improve both accessibility to computational resources, and the way bioinformatics algorithms are accessed and used together.

To improve accessibility to bioinformatics resources, several bioinformatics projects successfully aid researchers in accessing suites of algorithms (see Table 7). Well established applications like EMBOSS [433] and the SDSC Biology Workbench [434] provide access to a wide variety of algorithms. Additionally, an increasing number of new integrative projects have also become available for particular areas of research. For

gene regulation analysis, the Toucan workbench [259] and SeqVISTA [260] allow users to remotely access tools for comparative genomics, motif detection, and module detection. For expression profiling, integrative tools like Expression Profiler [435] and TIGR MeV [436] (MultiExperiment Viewer) incorporate algorithms for clustering, visualization, and statistical analysis.

Table 7: Integrated Bioinformatics Projects

GENERAL ALGORITHMS AND DATA	
	BioMOBY[437]
	LSID[438]
	BioPERL[439], BioJAVA[422]
	caBIG[440]
	eScience[441]
	BIAS[442]
	PathPort/Toolbus[443]
	GLAD[444]
	KDOM[445]
	Pise[446]
	EMBOSS[433], JEmboss[447]
	SDSC Biology Workbench[434]
	SeWeR[448]
	AnaBench[449]
	Celera Discovery System[450]
	myGRID[451]
PIPELINE	
	bioPipe[452]
	Taverna[453]
	Pegasys[454]
	Alfresco[455]
	FLOSYS[456]
	BOD[457]
GENE REGULATION	
	Toucan[458]
	Sockeye[403]
	SeqVISTA[260]
PROTEIN ANALYSIS/ GENE PREDICTION	
	exPASy[459]
	PipeAlign[460]

	DaliLite[461]
	GeneComber[462]
	GeneMachine[463]
	STING Suite[464]
	BioInfo3D[465]
	Molecular Biology Toolkit[466]
EVOLUTION/ COMPARATIVE GENOMICS	
	Taxonomy Workbench[467]
	PAL[468]
	SNAP[469]
	PLATCOM[470]
	Piptools[471]
EXPRESSION ANALYSIS	
	TM4[436]
	Expression Profiler[435]
	Biosphere[472]
SYSTEMS BIOLOGY	
	Systems Biology Workbench[473]
	Voyagene[474]
	KnowledgeEditor[475]
DATA MINING	
	WEKA[476]
	InfoEvolve[477]

A common constraint with each of these approaches is that they are dependent on one or more centralized resources; each approach requires prior knowledge of the relevant server locations on the Internet, either hard-coded into the application by developers or accessible through a service registry. Furthermore, the independence of each project may result in duplicated effort since many of these projects share similarities in their underlying bioinformatics toolsets. Our interest in integrating and comparing regulatory element discovery algorithms meant that we could create another suite for our particular analytical approach possibly overlapping integrative efforts of other suites in development; however, we reasoned that a more collaborative approach would improve

the accessibility and relevance of the analysis over a greater period of time. It is our hypothesis that a dynamic community-based approach to CPU usage, algorithm integration, and maintenance will improve the overall long-term quality of most integrative projects not just our own. By creating a means by which each project can take advantage of integrative efforts of the community, each individual project reduces the amount of maintenance that is required to provide reliable access to bioinformatics algorithms.

3.2 Methods

3.2.1 Chinook Architecture

We aimed to fulfill three requirements when implementing a peer-to-peer platform for bioinformatics analysis; the discovery of algorithms on the network and submission and retrieval of analysis should be available to users regardless of their operating system, the integration of new algorithms should be as simple as possible, and that the peer-to-peer component should have minimal consumption of CPU resources. To facilitate these requirements, the Chinook software is divided into three major components: the client node, the server node, and the peer-to-peer node (see Figure 7).

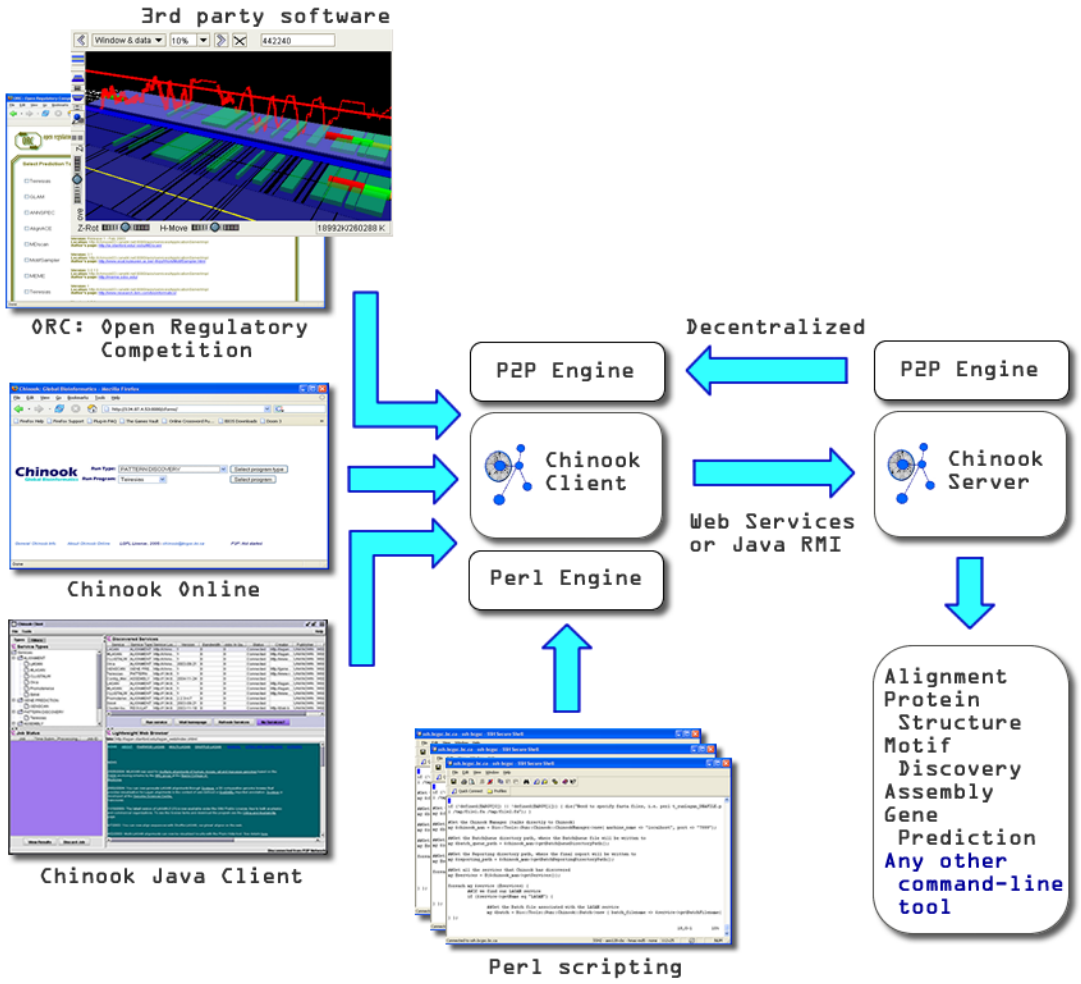

Figure 7: Chinook platform.

Distributed servers advertise command-line algorithms across the Internet. Discovered algorithms can be accessed from online, a graphical client, Perl scripting, or various other Chinook-integrated applications. The ubiquitous availability of algorithmic support through the Chinook Client to various applications reduces the amount of time application developers need to spend (re)integrating support for different algorithms. (Reprinted by permission from Macmillan Publishers Ltd: Nature Methods, Montgomery et al., 2(8): 563, copyright 2005.)

The client node is designed to facilitate algorithm discovery and job submission/retrieval either through a Java-based GUI, a Struts-based web application, or a BioPerl-based Perl module. Client nodes discover bioinformatics services by connecting to a running instance of the peer-to-peer node and retrieving a list of algorithms at different server node locations on a computer network. Once a desired

service has been selected, the client then determines which location to submit their analyses based on the response time of the server and the number of jobs being processed at the location. The client communicates directly with the server to process their analysis. When an analysis is submitted to a remote server, the client is able to monitor the standard output and standard error streams to observe the processing state of their job. On job completion, the client is able to download all the result files including any information written to the standard output or standard error stream during algorithm execution. Chinook only provides the downstream result files from any analysis; end users must make a choice as to how to visualize or subsequently process these files.

The peer-to-peer node is a Java application that facilitates discovery and advertisement of bioinformatics services over the network. The separation of the peer-to-peer node from the client and server nodes allows many clients or servers to exist as an individual network identity; for instance, a user could have one client and two servers active and all three instances of these components would be identified as belonging to a single network user. This configuration allows organizations to configure network presence for their services by specifying *n* peer-to-peer nodes that are accessible to users; thereby, reducing the individual CPU usage for discovery and advertisement operations.

The server node defines bioinformatics services using XML. XML service descriptions encode information about the parameters and default values that are accessible to a client; they also encode information about where target applications exist on the server and what modifications may be required to run them. The XML scheme in use is documented in the Chinook User Guide [478] but familiarity with the XML scheme is not required by as a server-side GUI is available to aid in customizing and

adding new services. By using XML to describe bioinformatics services, new algorithms can be integrated and deployed without requiring programming. A server node advertises their bioinformatics services by connecting to a running instance of the peer-to-peer node and providing the name, type, and network location of algorithms it hosts.

Chinook servers also provide access to configurable range of data sources. We have plugged access to the EnsEMBL database into Chinook to facilitate ease of sequence retrieval. Adding new data sources and their associated client data entry mechanisms is facilitated using Java reflection [479]. Client-server compatibility is maintained by restricting clients to only the data entry mechanisms that they currently possess.

3.2.2 Client-server communication

The peer-to-peer node utilizes the JXTA software library [480] to advertise and discover service locations across physical network boundaries. JXTA is an open source API designed to allow peer-to-peer communication between any networked devices (ranging from cell phones to desktop computers).

Client-server communication is facilitated using the Java Remote Method Invocation (RMI) protocol and/or Web Services. Servers using RMI require no special configuration except an open registry port to allow incoming TCP/IP requests. Servers using Web Services are deployed to a Tomcat Server running Apache Axis; we typically use the default 8080 port for servers running over Web Services.

3.2.3 Availability

The software and code is freely available from http://www.bcgsc.bc.ca/chinook/. Chinook components are deployed and run from ANT; the Chinook client comes with its own installer or can be run from Java Web Start. The Chinook web application is run on Tomcat and can be mirrored by downloading a WAR file. All Chinook software and ORC software is licensed under the Lesser GNU Public License. The minimum requirements for Chinook are a Java installation (1.4 or greater) and 256 MB of RAM. Tools currently available through Chinook are listed in Table 9.

Table 8: Bioinformatics tools in Chinook.

ClustalW[175]
Conreal [481]
DIALIGN[482]
LAGAN[176]
Mauve[483]
ORCA (Wasserman WW, unpublished)
Shuffle-LAGAN[484]
T-Coffee[485]
Promoterwise (Birney E, unpublished)
Primer3[486]
Eponine[91]
ANN-Spec[100]
ELPH [107]
Recursive Gibbs Motif Sampler[108]
MEME[487]
Motifsampler[104]
RSAT oligo analysis[488]
STUBB[489]
Teiresias[98]
wConsensus[490]
Genscan[491]
Sim4[492]
MSCAN[230]
Cluster-Buster[229]
Clover[493]
Weeder[101]
AlignACE[102]

MDScan[103] GLAM[99]

3.3 Results and Discussion

We have developed a decentralized peer-to-peer platform for the exchange of analysis utilities. Using the principles of common file-sharing applications, we enable users to advertise, discover, and execute utilities across a computer network without the requirement of a centralized server or administrative body. We have hypothesized that this type of technology will greatly reduce the parallel efforts of bioinformaticians by providing a shared foundation for their algorithm integration activities. Furthermore, in developing a peer-to-peer platform for the bioinformatics community, we were able to identify and address several challenges that are currently faced by end users of bioinformatics applications, namely wet-lab biologists.

3.3.1 Peer-to-peer for end users

A significant barrier to algorithm utilization by end users (laboratory biologists, bioinformaticians, and students) is that it is often difficult to identify the state-of-the-art. The diverse landscape of bioinformatics utilities, the complexity of their interfaces and underlying algorithms, and the competitive interests of rival groups make the problem appear untenable; especially when most end users want to quickly utilize and compare a combination of well-established and novel methods and then return to core tasks. A peer-to-peer system has the advantage of being adaptable to community standards and interests. As new utilities become available, they are instantly accessible to end users.

When these utilities become adopted and served at more locations, end users can choose to use them based on their higher presence in the network. As long as algorithm popularity is a reliable function of a particular algorithm's usefulness, end users will be able to discover and utilize these algorithms and compare their results to other popular utilities, whether they are simply different versions of the same algorithm or different algorithms altogether. Furthermore, as more locations become available, end users have access to an underlying computational resource that could facilitate application of the algorithm to large datasets previously regarded as too computationally intensive.

A peer-to-peer approach further satisfies subtle barriers to algorithm utilization; in many cases, end users cannot use particular algorithms due to incompatible installation requirements, lack of documentation, and/or lack of support. Our peer-to-peer system requires that utility maintenance is performed by utility providers instead of end users. This requirement is of significant advantage to end users as bioinformatics utilities can be accessed across operating systems, without requiring the end user to install dependencies, maintain versions, and possess specific hardware. To enable an end user to obtain more information about utilities they have discovered, we have integrated a web browser and author links into our system. This web browser allows them to visit associated documentation or the relevant homepage for any particular service; thereby providing a mode of attribution for the original authors of the utility and a method by which end users can investigate particular utilities in further detail. Furthermore, utility providers have the opportunity to publish their e-mail addresses to end users; this information provides a community of contacts for end users interested in performing a particular analysis in addition to the original author of the utility providing the support themselves. In many

cases, this mode of contact is preferable when end users want to discuss the particular benefits of various approaches among individuals who have utilized them.

As multiple barriers impede communication between the bioinformatics and end user communities; regardless of whether they are organizational or individual challenges, we have hypothesized that a collaborative network would aid in providing bioinformatics analysis to the end user community and thereby improve the accessibility of the current approaches as well as their long-term quality of application suites built on this architecture. A peer-to-peer system provides a mechanism for end users to discover and utilize new and well-established bioinformatics algorithms without requiring them to identify, install, and maintain the utilities themselves. It is our belief that the large scale adoption of an open peer-to-peer system by several bioinformatics groups would also allow end users to target various algorithms based on their cumulative presence throughout the network. We also believe that adoption of a peer-to-peer system at an institutional level can improve the accessibility of bioinformatics resources within individual organizations and consortia.

3.3.2 Peer-to-peer for bioinformaticians

The domain of bioinformatics development is often described as fractious; Lincoln Stein is oft-quoted for comparing the efforts of bioinformaticians to the feudal states of Italy [494]. The landscape of bioinformatics is changing though. As described in Table 7, many integrative projects aim to improve the accessibility of biological data and bioinformatics analysis to end users. We propose that a peer-to-peer platform for analysis integration could improve the accessibility of algorithms for each of these

projects and potentially provide bioinformaticians with unprecedented computational resources. An open community-based approach to algorithm integration allows each participant to benefit from the integration efforts of the entire group. However, a significant barrier to adoption by the bioinformatics community to a common integration infrastructure is correlated to the ease in which new utilities are integrated. We have developed a XML schema as part of our system to enable users to quickly describe and integrate new utilities; new utilities can also be added through a graphical user interface. By making server integration as easy as possible, our peer-to-peer system encourages the advertisement of redundant utilities. This functionality not only provides a method by which end users can determine which algorithms are commonly-used across the community, but it also helps to both facilitate distributed computing applications, and prevent the failure of centralized registries or servers, as no individual service location is essential to the maintenance of the network.

In creating a peer-to-peer platform for bioinformatics analysis, we had to design to specifically meet the needs of the academic bioinformatics community. Our platform will allow new algorithms to get rapid exposure in the community but, as service providers adopt these algorithms, also provide mechanisms for proper attribution. This was particularly well-iterated by DJ States in response to Stein's call for an integrated bioinformatics community where he stated, "[interoperable systems are] in direct conflict with the need for recognition and citation and will do nothing for the career of the developer" [495]. For this reason, we designed our peer-to-peer system with scientific attribution in mind. A web browser is integrated into our system to take users directly to the web sites of original authors (whether this points to a peer-reviewed manuscript or

their home web site). Furthermore, all service information is accompanied by information about the original service developers. By paying particular attention to these issues, we aim to have original developers take an active interest in providing their services through this system.

This type of technology is further amenable to administrators looking for an easy way to make their high-performance compute resources available locally or worldwide. Bioinformatics labs can easily introduce sequence or annotation data into our system allowing them to customize their nodes for specific types of in-house data. Our usage of cross-platform technology like Java and Web Services, and our integration with Perl for batching jobs over our system, makes usage compliant with standard bioinformatics practices.

Chinook is currently being integrated and/or used in the Sockeye [403] and BioMoby [437] applications. Each application is developed by a different lab and provides different types of analysis for several bioinformatics algorithms. By sharing a common integration framework we have created a small collaborative network in our community, where in many cases bioinformatics algorithms are maintained directly from the original authors. Furthermore, Chinook also integrates the EnsEMBL [496] and Jaspar [497] databases to facilitate specialized data entry. Currently, servers are available at the BC Genome Sciences Centre, the Centre for Molecular Medicine and Therapeutics, CANARIE Inc., and at Boston University. We plan to extend this system to more labs in our community with an open invitation for other labs to provide their own services. We provide support for this activity at chinook@bcgsc.bc.ca and http://www.bcgsc.bc.ca/chinook.

3.3.2 Application of peer-to-peer approach to assessment of computational tools for transcription factor binding site discovery

To demonstrate the applicability of our method, we have applied our peer-to-peer software to the assessment of computational tools involved in transcription factor binding site discovery. This activity is particularly amenable to a peer-to-peer approach as new algorithms for discovery of transcription factor binding sites become increasingly available.

A previous analysis of transcription factor binding site discovery tools involved sending datasets to multiple tool developers and then assessing their results against undisclosed positive controls [97]. This allowed tool developers to apply their algorithms using desired parameter settings thereby creating a benchmark against which other tools could be assessed. However, this analysis uses static datasets from four organisms and fifty-four control datasets, and requires coordinating the analyses of multiple tool developers. We have used the Chinook peer-to-peer platform to create a web application called the Open Regulatory Competition (ORC) which performs immediate sensitivity, specificity and positive predictive value measurements for a dynamically discovered set of transcription factor binding site discovery programs across the Internet. Furthermore, the correlation coefficient for all pairs of tools is generated and reported (Figure 8). A user of ORC is able to quickly analyze new algorithms as they come online against existing algorithms for optimal performance against their own datasets. This utility is especially relevant as new datasets applying novel genomes and identified transcription factor binding sites become available.

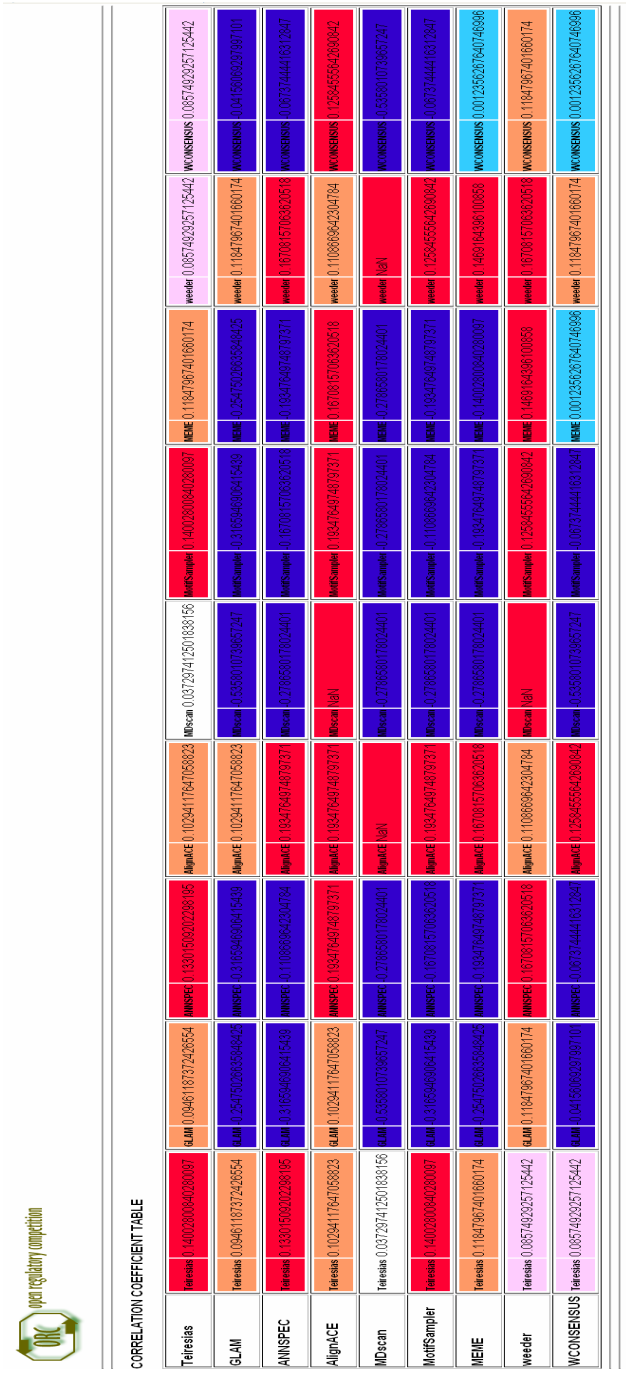

Figure 8: Open Regulatory Competition Correlation Coefficient Table

Results from several transcription factor binding site discovery programs are compared using random data. The correlation coefficient statistic is a measure of correlation between known and predicted sites. A correlation coefficient of -1 indicates perfect anti-correlation whereas a value of +1 indicates perfect correlation. This chart shows correlation coefficient values for random data (performance ability should not be assessed from this plot) for all combinations of ORC-integrated tools

This approach demonstrates the use of our platform in facilitating ongoing competition and assessment of a wide-range of types of computational tools against novel datasets. This system is further amenable to evaluating alignment, gene prediction, or regulatory module prediction in the near future.

ORC is packaged as a WAR file for easy mirroring, is open-source to support addition of new types of analysis, and has a Wiki page to support discussion of results.

3.4 Conclusions

This work demonstrates the use of peer-to-peer technology to organize and synthesize diverse bioinformatics tools. As part of this thesis, I was particularly interested in how this technology might avail itself to identifying and comparing alignment and motif discovery algorithms. The ORC application represents a strategy that is tractable to these needs – particularly the open competition of established or novel datasets through a web interface. This strategy allows any researcher, independent of their background, to quickly evaluate the state-of-the-art in the field. The peer-to-peer component further introduces a novel selection pressure mechanism to a burgeoning population of bioinformatics resources. Through this work, I became acutely aware of how limited existing regulatory element datasets are and how no substantial repository of functionally regulatory polymorphisms exists. The focus of the next chapter is a strategy that I developed to assemble a database that supports annotation of regulatory information in perpetuity by the researcher community and the results of months of focused literature review to assemble a high-quality set of functional regulatory polymorphisms that could be used for the analyses as part of Chapter 5.

Chapter 4: ORegAnno: An open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation.

A version of this chapter has been published:

Montgomery, S.B.[1,], Griffith, O.L.[1,], Sleumer, M.C., Bergman, C.M., Bilenky, M., Pleasance, E.D., Prychyna, Y., Zhang, X., Jones, S.J.M. 2006. An open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics.* **22(5)**:637-40

Co-authorship details: I was responsible for initial design of the ORegAnno resource and all computational implementation of this resource. The majority of co-authors listed in the original publication have added valuable data to this resource. Of particular note, Obi Griffith managed the population of this resource including several regulatory SNPs, added the evidence ontology, and helped improve the schema and design as the database developed. I wrote the published manuscript.

4.1 Introduction

The effectiveness of bioinformatics methods for identifying regulatory regions in genomic sequence is dependent on our understanding of gene regulation biology in its natural state. This is particularly evident in that models of transcription factor binding in regulatory regions have underpinned the development of such bioinformatics methods as phylogenetic footprinting, transcription factor binding matrices and motif clustering (reviewed in [73]). However, the predictive ability of algorithms which implement these methods has been predominantly indeterminate, as their assessment has relied on datasets containing few biologically-validated regulatory regions [97]. To enrich these datasets, several databases have been designed to independently organize the sites of promoter activity [115,117,123,125-127] transcription factor binding [114,129,130,132] and regulatory variation [328,335,389]. Several challenges face the user when accessing these databases for the annotation of biologically-validated regulatory regions. For many databases, considerable investigation can be required to collate its information, determine the original experimental techniques used, determine the "genomic scope" of the annotation (i.e. what further annotation is in the vicinity and informative), obtain a sequence of sufficient length to map to new genome sequence assemblies, cross-reference or follow-up on specific annotation or access the annotation programmatically. Furthermore, as new regulatory sequences become characterized each database requires its own curators *ad infinitum* as few or no mechanisms currently exist in which a community of researchers can add to or comment on these annotations. The Open Regulatory Annotation (ORegAnno) database has been developed to address these issues

and provide a unique platform for community annotation of experimentally verified regulatory regions.

4.2 Description of the ORegAnno database

ORegAnno permits open annotation of regulatory regions by providing roles and secure user accounts to contributors. Three roles exist for ORegAnno contributors: user, validator and administrator. A user role enables a contributor to add individual annotations of promoters, transcription factor binding sites and regulatory mutations to the database. As a first step in validating a new annotation's authenticity, each submitted annotation is immediately cross-referenced against PubMed [498], Entrez Gene [499], dbSNP [327], the NCBI Taxonomy database [498], and EnsEMBL [21]. Once submitted, the record is added to the database and an email is generated containing an XML representation of this record to members of the ORegAnno developers' mailing-list (oreganno-guts@bcgsc.ca). As a second step in validating an annotation's authenticity, a validator role enables a contributor to score individual annotations in the database. Validators will modify an overall score for an annotation based on their ability to confirm the reliability of annotation from literature. Validators have the option of increasing the annotation score by one if they can confirm the record, leaving the score unchanged if their conclusions are indeterminate, or decreasing the score by one if an error has been found. Each observation and score modification of an annotation along with the associated validator user information is stored in ORegAnno. An administrator role enables a contributor to assign roles, add or define evidence (classes, types, and subtypes) and batch upload large sets of annotations directly to the database. Both administrator

and validator roles allow the modification of records; for a record modification, a new record is created and the old record is marked as being deprecated by the newer record. Each role is further permitted to add comments to individual annotations to improve subsequent users' understanding of a particular annotation. ORegAnno's usage of roles provides a level of accountability in the database as users become owners of their annotation and validators become responsible for verifying an annotation's authenticity.

For each type of annotation that is currently in ORegAnno, the database obeys the following rules:

1) Each annotation describes a regulatory property of one target gene which is either user-defined, in Entrez Gene or in EnsEMBL.
2) Each annotation must be attributed to a species which has a taxonomy id in the NCBI Taxonomy database.
3) Each annotation can optionally be associated to a specific dataset. This functionality allows external curators to manage particular sets of annotation using the ORegAnno's curation tools.
4) Each annotation specifies an evidence type, subtype and class describing the biological technique cited to discover the regulatory sequence. Evidence classes are broken into two categories: the 'regulator' classes describe evidence for the specific protein(s) that bind a site. The 'regulatory site' classes describe evidence for the function of a regulatory sequence itself. These two categories are further divided into three levels of regulation (transcription, transcript stability and translation). Thus, a total of six evidence classes currently exist. Evidence types describe the generic assay

used while subtypes define specific implementations of these assays (Table 2). Each annotation can have multiple entries from any evidence class, type and subtype describing each piece of experimental evidence for the regulatory sequence and/or binding protein.

Table 9: Evidence types and sub-types

Evidence type	Evidence subtype
Electrophoretic Mobility Shift Assay (EMSA)	Direct gel shift
	Supershift
	Gel shift competition
Reporter Gene Assay	Transient transfection luciferase assay
	Chloramphenicol acetyltransferase (CAT) Assay
	In-vivo GFP Expression Assay
	Dual luciferase reporter gene assay
	In-vivo LacZ Expression Assay
Protein Binding Assay	Chromatin immunoprecipitation (ChIP)
	DNase Footprinting Assay
	Yeast 1-hybrid assay
RNA Expression Assay	RNase Protection Assay (RPA)
	Reverse Transcriptase Polymerase Chain Reaction (RT-PCR)
	Allele-specific Transcript Quantification (ASTQ)
	Competitive PCR (cPCR)
	RNA Ligase-mediated Rapid Amplification of cDNA ends (RLM_RACE)
	Whole-mount in situ hybridization
Protein Expression Assay	Western Blot Assay
	Enzyme-linked Immunosorbent Assay (ELISA)
	Luciferase Expression Assay
	Indirect Immunofluorescence
RNA Stability Assay	RNA synthesis blocking
SNP Discovery and Genotyping	Resequencing
	Single-Stranded Conformational Polymorphism (SSCP)
	Restriction Fragment Length Polymorphism (RFLP) Analysis
Orthologous gene conservation	Conservation found by alignment
	Conservation found by scanning with a motif model
Gene Co-Expression	Co-expressed genes determined through reporter gene experiments
	Co-expressed genes determined through microarray experiments
	Co-expressed genes determined through expression pattern

5) Each piece of experimental evidence is optionally associated to a specific cell type using the eVOC cell type ontology [204].
6) Each transcription factor binding site or regulatory mutation must specify a target transcription factor which is either user-defined, in Entrez Gene or in EnsEMBL. If there is no recorded gene target, a classification of "unknown" is specified.
7) Each transcription factor binding site or regulatory mutation must include sequence with at least 40 bases of flanking genomic sequence to allow the site to be mapped to any release of an associated genome.
8) Where available, any annotation can provide search space information specifying the region that was assayed, not just the regulatory sequence.
9) User information is recorded with each annotation.
10) Each annotation must reference a valid PubMed article. To reduce the entry of redundant annotations, a warning is issued if an annotation is found with either an existing reference identifier or matching genomic sequence.
11) For regulatory mutations, each variant or haplotype that has been proven to cause a change in gene expression is a separate record. The sequences containing both the wild-type and mutant sequences must be specified. If available, a dbSNP cross-reference can also be specified. The type of variant is specified as either being germline, somatic or artificial.
12) Each record is associated to a positive, neutral or negative outcome based on the experimental results from the primary reference. For instance, a sequence that was demonstrated not to bind a particular transcription factor could be

annotated as a negative outcome; however, to be meaningful, the associated evidence must provide adequate information to determine the conditions assayed.

ORegAnno comes equipped with analysis tools to assist in annotation of new records. In many cases, extracting genome sequence from literature and identifying the corresponding sequences in genome databases is problematic . ORegAnno provides the tools ENSSCAN for finding one or more specific sequences within distances relative to the start of an EnsEMBL transcript, ENSFETCH for retrieving small sequences within distances relative to the start of an EnsEMBL transcript (i.e. from -34 to -40 of the transcription start site), NCBISCAN for finding one or more specific sequences within defined distances of a GenBank-reference sequence, and NCBIFETCH for highlighting small (gapped) sequences within a GenBank-reference sequence.

4.3 Current content of the ORegAnno database

As of July 2006, the ORegAnno database housed a total of 3042 entries from over 100 users. These include 874 regulatory regions, 1991 transcription factor binding sites, and 177 regulatory mutations (polymorphisms and haplotypes) from 11 species (see Table 9). A large fraction of these sites were obtained from previous large-scale collections such as the FlyReg resource [114] and a large set of muscle/liver-specific regulatory sites curated by Wasserman, Fickett and others [197,220]. 11 regulatory polymorphism records were obtained from rSNP_DB [500]; rSNP_DB records were filtered to include only those records which pertained to natural mutations or polymorphisms. In addition, over 400 new annotations were obtained by manual curation

of literature. Thus, the ORegAnno resource represents an assembly of existing records, a significant addition of new records and provides an open-access system for continued, community based accumulation of sites within a standardized framework.

Table 10: Current contents of the ORegAnno database.

	Regulatory Haplotype	Regulatory Polymorphism	Regulatory Region	Transcription Factor Binding Site
Caenorhabditis briggsae	0	0	0	24
Caenorhabditis elegans	0	0	13	189
Danio rerio	0	0	2	0
Drosophila melanogaster	0	0	0	1350
Gallus gallus	0	0	5	22
Homo sapiens	6	170	781	233
Mus musculus	1	0	10	107
Rattus norvegicus	0	0	9	58
Saccromyces cerevisae	0	0	1	7
Xenopus tropicalis	0	0	0	1
Totals	**7**	**170**	**821**	**1991**

4.4 ORegAnno Publication Queue

The ORegAnno database has recently been upgraded with a publication queue. ORegAnno's publication queue allows registered users to input relevant papers from scientific journals to a queue system for annotation. All that is required is a valid PubMED ID (PMID). Each added paper is set to PENDING. Any user can explicitly OPEN publications from the queue that are PENDING, to begin the annotation process. Once a paper has been completely annotated, the user will set the publication state to CLOSED. Otherwise, the user can revert the state back to PENDING. Comment fields are available for each change of state in the queue. A publication must be in the work queue before it can be annotated. VALIDATORS can set the state of a CLOSED paper back to PENDING, if they feel something was missed, otherwise they may correct the annotation directly. The value of this system is that it is trivial for users to contribute to

the ORegAnno database and it is possible to capture a measure of the number of papers needed to be annotated in the scientific domain. Currently, the publication queue contains 286 PENDING publications, 7 OPEN publications and 588 CLOSED publications.

4.5 Access

The raw ORegAnno data is available directly over MySQL from db01.bcgsc.ca or through Web services [501]. Methods are exported using Web services to search for annotation by various fields enabling fetches by such fields as stable id, species, gene name, transcription factor name or cross-reference sources. ORegAnno also automatically maps each annotation to its relevant genome using Blast [267]; these mappings are viewable through the UCSC Genome Browser [20] or EnsEMBL using the Distributed Annotation System [502]. Finally, the entire database is converted to XML format and made available on the website daily. The ORegAnno web application is open-source under the Lesser GNU Public Licence thereby permitting all forms of modification and mirroring.

4.6 Conclusions

The ORegAnno resource represents the first open-access, community-based forum for annotation of regulatory sequences. ORegAnno is currently the largest collection of functionally-validated regulatory annotations available with unrestricted access. To our knowledge, it is the first resource to incorporate regulatory regions, binding sites and variation into a single resource. It is also the first system to incorporate

a structured system for experimental evidence and allow both negative and positive results. The requirements for sufficient flanking sequence and verified gene identifiers (Ensembl or Entrez) ensure maximum compatibility with the community's various research needs, both currently and in the future. The intention of ORegAnno is not to replace any regulatory element databases. Many of the well-targeted databases have domain- or species-specific information that would be impractical to incorporate into a single resource. Instead, we hope to create a single multi-species database and curation system for some of the most essential information (target gene, binding protein, binding site sequence, etc.). Thus, we believe ORegAnno should exist in collaboration with the more specific databases as a central warehouse of data, with the ultimate goal of incorporating all experimentally-verified regulatory annotation. We anticipate that this growing library of regulatory elements will prove an important resource for the validation of computational methods of motif detection, investigations of regulatory element evolution and an essential resource for the appraisal and validation of genome-wide regulatory predictions [264,268].

Of the 125 users that have signed up for ORegAnno accounts, 12 have entered ORegAnno records. However, since the publication of this work, an international collaboration has assembled to help populate papers into the queue using text-mining strategies and has developed the RegCreative jamboree with the specific aim of populating the ORegAnno resource [503]. My focus, as part of this thesis, since development of the ORegAnno resource is to continue to maintain it and to use the regulatory polymorphisms I have curated within ORegAnno to begin training approaches

for prioritizing variants in the promoter regions of genes; it is this work that is the focus of Chapter 5.

Chapter 5: A computational discrimination strategy for regulatory polymorphisms in the promoter regions of *Homo sapiens*.

5.1 Introduction

Our ability to identify the molecular mechanisms responsible for specific genetic traits within our population will be enhanced by our imminent ability to decipher each individual's genome. This is evident from recent advances in sequencing and genotyping technologies, which allow an increasing number of variants to be sampled for association and linkage (reviewed in [504-506]) and contribute a growing number of sources of variation and their frequencies to public databases each year. As new variants become identified, each becomes a molecular window into our past, present, and future--each aids in tracing our genetic heritage, helping to chart the footsteps of our common evolution and possesses the potential to predict disease or drug susceptibilities, acting as an early-warning system in initiating preventative medical practice (reviewed in [507,508]). However, our ability to catalogue genotypes has far outstripped our ability to associate them with phenotypes. Currently, over 6 million single-nucleotide polymorphisms (SNPs; from a pool of over 27 million submitted SNPs) exist in version 126 of dbSNP [327]; of these SNPs, only a very small fraction have been associated to a phenotype using genetic association or linkage analysis assays. The reasons for this are because these assays are costly, time-consuming, and dependent on the frequency of the genotype in the sampled population. To select candidates for functional validation, computational methods have been developed to identify SNPs that alter the protein-coding structure of genes [337,359-367]. These types of computational methods predominantly prioritize potentially harmful SNPs by identifying those SNPs which alter a protein's amino acid sequence, are targeted at well-conserved regions or functional protein domains, and alter the biochemical structure of the protein. However, very few methods identify those

SNPs which alter the expression of genes. Such SNPs have been implicated in the etiology of several human diseases, including cancer [397,398], depression [399], systematic lupus erythematosus [400], perinatal HIV-1 transmission [401], and response to type 1 interferons [402]. This work has aimed to extend computer-based techniques to identify this particular class of functional variants within the core promoter region of human genes.

Conventional computational approaches to regulatory SNP classification have predominantly relied on allele-specific differences in the scoring of transcription factor weight matrices as supplied from databases like TRANSFAC and Jaspar [366,367,388]. SNPs which are located within matrix positions possessing high information content are assumed more likely to be functional. Support for this hypothesis to-date, however, has been restricted to single case examples and small test sets. Furthermore, a recent study has failed to detect significant weight matrix signals in more than 33% of regulatory polymorphisms [509]. The prevailing hypothesis in computational regulatory element prediction has been that the majority of predictions using unrestricted application of matrix-based approaches are false positives. By extending this technique and using phylogenetic footprinting between mouse and human, it was demonstrated that from 10 SNPs which show significant allele-specific differences in Jasper predictions, 7 also demonstrated electrophoretic mobility shift differences [388]. However, only 2 of the 7 had marked effect in reporter gene assays. Conservation alone has also been demonstrated as a poor discriminant of function in a study of regulatory polymorphisms in EPD promoters where 0 of 10 experimentally-validated regulatory variants were in conserved binding sites [136].

A substantial challenge with developing strategies for identifying functional non-coding variants has been the shortage of characterized regulatory variants. Most studies stop once a susceptibility haplotype has been ascertained. While knowledge of susceptibility haplotypes alone will enable characterization of an individual's genetic risk predispositions, the development of screening techniques, which identify the molecular components altered by genetic variants, will be necessary in the development of targeted pharmacological treatments.

To address this problem, we have assembled the largest openly-available collection of functional regulatory polymorphisms within the ORegAnno database. From this dataset, we have then looked at several features of these SNPs as they relate to polymorphisms of unknown function (ufSNPs) within the promoter regions of associated genes (up to 6kb). Our hypothesis has been that from an unknown combination of regulatory and population genetic properties, the discriminative efficacy of individual properties can be evaluated and significant predictors of function can be chosen. Within our assayed set, we have found that the single biggest discriminants are the distance to transcription start site, local repetitive density and content, minor and derived allele frequency, CpG island presence, DNaseI hypersensitive site presence, and sequence conservation. Notably, the unrestricted application of a matrix-based approach is demonstrated to be one of the least effective classifiers.

We have used this dataset to train a support vector machine (SVM) classifier. Two approaches were used to train the classifier: one, where the properties of all positive SNPs were compared to all ufSNPs and the other where each of the positive SNPs and ufSNPs within an associated gene are compared to the average values for each property

within that gene (termed here the "ALL" and "GROUP" approach, respectively). The "ALL" approach is designed to ascertain if there are any discriminants that are important across the test set, while the "GROUP" approach is designed to ascertain if there are important directional shifts in values within a gene that may discriminate functional SNPs from ufSNPs. In a cross validated test, the SVM achieves a ROC value of 0.89 +/- 0.06 (ALL; Sensitivity 0.85 +/- 0.12; Specificity 0.80 +/- 0.06) and 0.78 +/- 0.07 (GROUP; Sensitivity 0.74 +/- 0.19; Specificity 0.65 +/- 0.14). We have used the SVM to prioritize polymorphisms of importance to cancer.

5.2 Methods

5.2.1 Data

Literature describing genetic regulatory polymorphisms with direct effects on gene expression was manually curated from PubMED [498]. 171 relevant polymorphisms were identified in 106 publications. Each polymorphism was strictly selected based on supporting evidence which identified them as causal variants and not simply in linkage disequilibrium with a functional polymorphism. From this set, 109 polymorphisms were selected as containing only SNPs (indels were excluded) and being within 6kb of the transcription start site of their associated gene (as annotated in version 37 of EnsEMBL [21]). In total, this 109-rSNP set contained polymorphisms involved in altering the expression of 88 different transcripts. Figure 9 illustrates the types of original experiments that were performed within the 109-rSNP set and Figure 10 illustrates how these experiments were used to validate the effects of individual SNPs.

Using each of the 88 transcripts, SNPs within 6kb of the transcription start site were extracted from version 37 of EnsEMBL (dbSNP version 125)--2690 SNPs of unknown function (ufSNPs) were obtained. The ufSNPs and 109-rSNPs have been mapped and are available as supplementary material at http://smweb.bcgsc.ca/thesis/supplementary.html.

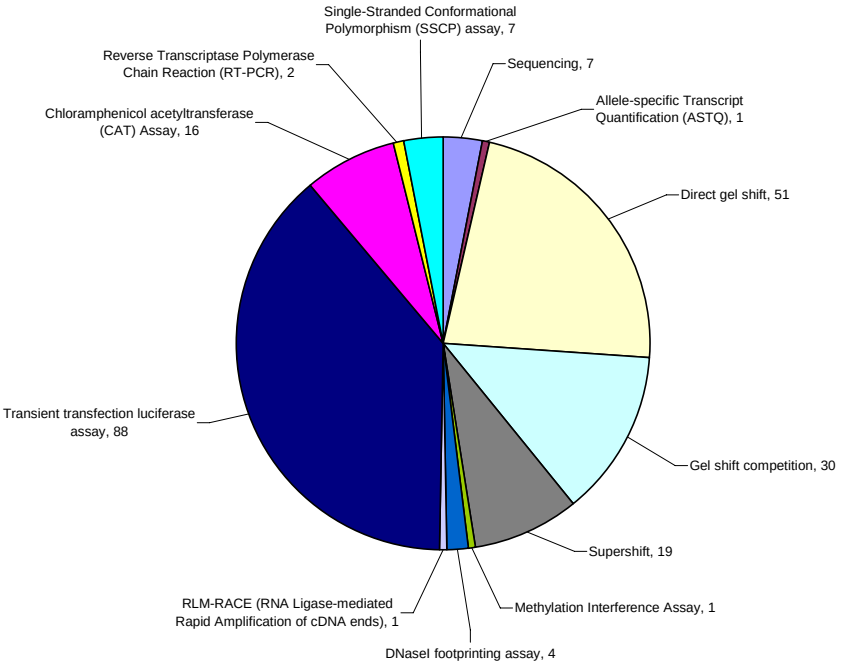

Figure 9: Supporting evidence for regulatory SNP identification (by experiment)

Experimental assays conducted over the 109-rSNP set. The majority of experiments were gel shifts and luciferase assays.

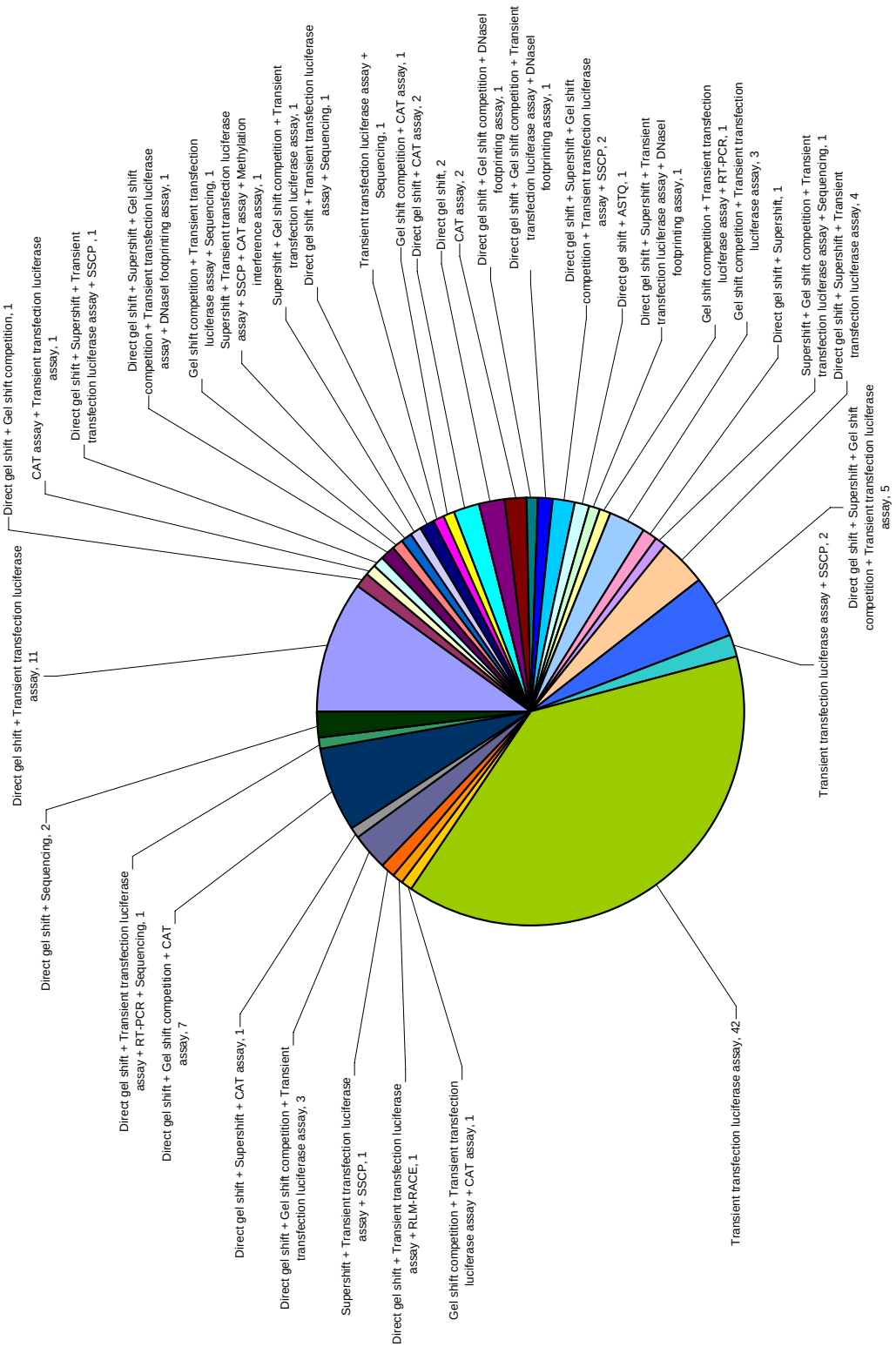

Figure 10: Supporting evidence for regulatory SNP identification (by record)

Experimental assays conducted over for each of the 109-rSNPs. The majority of rSNPs were confirmed using transient transfection assays alone.

5.2.2 Investigated Properties

Twenty-three different properties were assessed for each SNP in both the 109-rSNP and ufSNP sets. Each is discussed here.

TRANSFAC ANALYSIS (Property 1)

A differential Transfac analysis (1) is performed by substituting the canonical nucleotide with its associated variant allele and measuring the difference in predicted binding site scores for the cumulative set of predicted factors.

$$\Delta s = \sum_{factor=i,j}^{i+j} (score_{(canonical)factor} - score_{(\text{var } iant)\ factor})$$

Here, Δs is the cumulative difference in the predicted binding site scores between the set of predicted factors, i, from the reference sequence and the set of predicted factors from the variant sequence, j.

MOTIF DISCOVERY ANALYSIS (Properties 2, 3, 4 and 5)

For each selected SNP, a 1kb nucleic acid sequence was retrieved from EnsEMBL (NCBI35). The EnsEMBL compara database was subsequently utilized to retrieve pre-calculated orthologous sequences from completed genomes; specifically, sequences from chimpanzee, rhesus macaque, mouse, dog, rat and chicken were used. A differential Weeder [101] and MotifSampler [104] analysis was performed by separately inputting canonical and variant human sequences with the set of associated orthologues and recording both the predicted score and the magnitude of difference between scores predicted for motifs overlapping the tested polymorphism (this is used to measure both how a allele-specific change effects scoring and the strength of the motif predictions). To

improve the probability of detecting the desired motif, Weeder was set to detect 500 motifs and MotifSampler was seeded with 25bp around the polymorphism. These tools were selected both because of their different approaches to motif discovery (Weeder is enumerative and MotifSampler is based on optimizing an objective function) and because they have been previously demonstrated to have moderately complementary performance characteristics [97]. A 1kb region was selected to allow duplicated motifs to contribute to the scoring function and to permit relax positional constraint on contributing motif position.

DNA CURVATURE (Properties 6 and 7)

A differential DNA bendability (4) and curvature (5) analysis is performed on canonical and variant sequences using an implementation of the BEND algorithm called "banana" and packaged in the EMBOSS toolkit [433,510]. The effects of DNA structure in mammalian systems remains largely unascertained, previous characterization in bacterial systems has demonstrated its role in creating conditions suitable for transcription factor binding [240,241].

LOCAL GC CONTENT AND DNA THERMODYNAMICS (Properties 8 and 9)

The differential effects on local GC content (6) and thermodynamic stability of the DNA sequence (7) are assessed using the "dan" application packaged in the EMBOSS analysis package [433]. The presence of functional transcription factor binding sites in GC-rich sequences has been previously studied [82,90].

DISTANCE TO TRANSCRIPTION START SITE (Properties 10 and 11)

The distance to the transcription start site (TSS), as annotated by EnsEMBL, was recorded. Distance to TSS has been previously identified as a significant discriminant of regulatory polymorphisms; a study of 674 haplotypes in 247 gene promoters identified that sequence variants altering expression by 1.5-fold or more are preferentially located within the first 100-base pairs [509]. Both the raw distance and the logarithm of the distance were used.

LOCAL REPETITIVE CONTENT (Properties 12, 13, 14 and 15)

Local repetitive content of a 200-bp DNA segment centred on the assayed polymorphism was calculated using repetitive annotation curated in EnsEMBL. Four different metrics were assessed in this region: 1) the percentage of repetitive bases; 2) whether the polymorphism was in a repeat or not; 3) the number of repeats of length greater than 1kb; and 4) length less than 1kb that overlap this region. Each value was normalized to its expectancy at the calculated distance from the transcription start site in the associated chromosome (see 5.2.6).

MINOR AND DERIVED ALLELE FREQUENCIES (Properties 16 and 17)

Minor allele frequencies were obtained from dbSNP (version 125) directly using the “eutils” service. Each allele frequency was calculated from combining frequencies in all assayed populations. Derived allele frequencies were calculated by aligning a 1-kb human region centred on the polymorphism with orthologous chimpanzee sequence using ClustalW to obtain the derived allele and matching this allele with previously calculated

allele frequencies. Of the 109-rSNP and ufSNP sets 82 and 1178 had genotype data, respectively.

OPOSSUM/JASPAR COEXPESSION ANALYSIS (Property 18)

oPOSSUM was run to short-list a set of transcription factor binding matrices for differential analysis (as in the TRANSFAC test) [197]. Coexpression data was extracted from the TMM coexpression set published by Pavlidis et al. [511]. Coexpressed genes were broadly-selected based on at least one study reporting coexpression.

CPG ISLAND (Property 19)

CpG islands were obtained from annotation in the UCSC genome browser [20]. Whether or not a polymorphism was in a CpG island was recorded. This value was normalized to its expectancy at the calculated distance from the transcription start site in the associated chromosome (see 5.2.6).

DNASEI HYPERSENSITIVE SITES (Properties 20)

DNaseI hypersensitive sites were obtained from predictions as per Noble et al [512]. The latter set was mapped from hg15 to hg17 coordinates. Whether or not a polymorphism was in a DNaseI hypersensitive site was recorded. These values were normalized to its expectancy at the calculated distance from the transcription start site in the associated chromosome (see 5.2.6).

CONSERVATION USING PHASTCONS AND REGULATORY POTENTIAL (Property 21 and 22)

Conservation scores from both the PhastCons [162] and Regulatory Potential [163] methods were obtained from the UCSC genome browser. The local conservation of the polymorphism, as calculated by these scores, was recorded. These values were normalized to their expectancy at the calculated distance from the transcription start site in the associated chromosome (see 5.2.6)

CLUSTALW ALIGNMENT DEPTH (Property 23)

Each orthologous sequence set for an individual polymorphism was aligned using ClustalW [513] and the total evolutionary distance was calculated from the generated phylogenetic tree. This value measures the level of conservation and sequence divergence around each polymorphism.

5.2.3 Test design (ALL and GROUP)

Two types of data sets were assembled using the 23 properties ascertained for each SNP. One, an all-vs-all dataset, where the values of the 109-rSNP and ufSNP set are compared *en masse*. The other, a group analysis, where the average value of each property within an upstream region was calculated and individual SNP properties were recalculated from their divergence from this average. The ALL test is designed to identify global characteristics of rSNPs while the GROUP test is designed to look for directional trends within transcripts which might be indicative of a difference in function.

5.2.4 Support Vector Machine

Calculated data was input to the Gist SVM implementation [514]. Gist was run using the default parameters. This approach has been previously described in detail elsewhere [512]. Of note, the Gist SVM requires that every value in the test and training parameter space is not empty. To address this, where appropriate, properties are calculated as an allele-specific difference and the null hypothesis dictates that this difference should be as close to zero as possible. The ALL SVM was filled with gene specific average values wherever data could not be calculated. The GROUP SVM was filled with zero-values wherever data could not be calculated, indicating no divergence from average within the GROUP test set.

5.2.5 Performance Measurement

The individual importance of each property in discriminating regulatory polymorphisms was assessed in the ALL and GROUP test sets using a two-sampled t-test.

The performance of the Gist SVM classifier is measured using a receiver operating characteristic (ROC) curve. ROC scores of 1 indicated perfect discrimination, while those nearer to 0.5 indicate random classification of the input SNPs. ROC performance measurements have been previously described in detail elsewhere [512].

A ten-fold cross validation was performed to assess the overall performance of the SVM. 10% of the transcripts (and their associated SNPs) were randomly excluded and the remaining 90% was trained on. This analysis was performed ten times to calculate an average ROC value for the SVM.

SNPs in the promoters of several cancer genes were assessed using the GROUP SVM with and without distance to TSS properties included. A high-ranking list of SNPs from both these SVM classifications was manually-selected by identifying SNPs that scored highly in both classifications.

5.2.6 Distance normalization

We were concerned that several properties may be indirect measurements of distance from the TSS and that any discrimination strategy would be limited to characterizing this property alone. This concern is a particular challenge since distance ascertainment bias exists; most SNPs surveyed were within a few hundred basepairs of the TSS which is much smaller when compared to our sampling distance of 6kb. Furthermore, it has been well-established in a previous study that distance to TSS is correlated to detection of regulatory polymorphisms (it is unknown if this is because they are more likely to effect essential transcription factor binding sites or because there are a higher density of TFBS in these regions) [509]. For this reason, the discrimination potential of distance to TSS could not be ignored. To adjust for bias, however, we calculated the expectancy of observing a feature at a particular distance from the TSS for each individual chromosome (Figure 11). This expectancy value was used to normalize the observation values for several of the properties in this study. The impact of this normalization can be seen when comparing normalized GROUP ROC values against unnormalized GROUP ROC values; using a ten-fold cross validation, the unnormalized ROC values are 0.79 +/- 0.05.

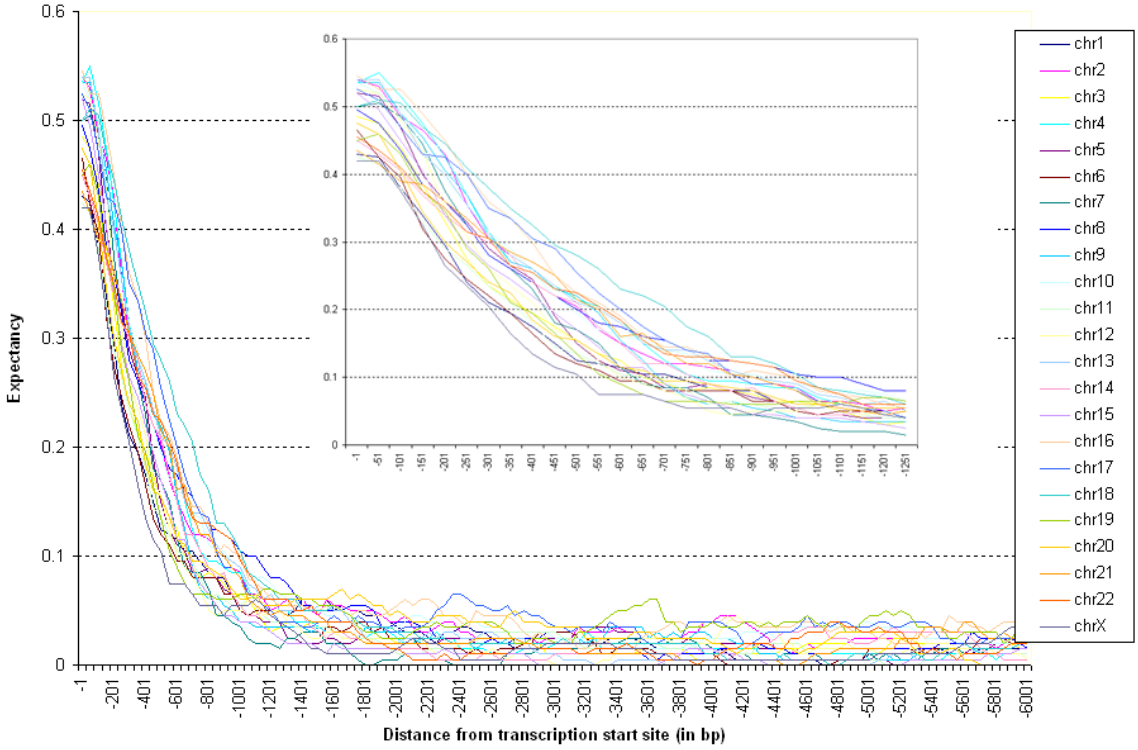

Figure 11: CpG island positional bias.

Cpg island expectancy is plotted for each chromosome as a function of distance from the transcription start site. This type of data was used to normalize many of the features in this study for distance from TSS. In this figure, the expectancy of being in a CpG island at position -1 for any promoter region is ~0.5. This suggests a mixture of promoter types which could drive the CpG island results lower through normalization in this study. It has been shown that without this normalization, higher CpG island values are a significant discriminator (data not shown).

5.3 Results

5.3.1 Discriminant classification

109 regulatory SNPs (109-rSNPs) and 2690 SNPs of unknown function (ufSNPs) in the promoter regions of 88 different genes were compiled to test properties that may discriminate polymorphisms with effects on gene expression. A two-sampled t-test comparing the ALL test set identified several properties of significance in discriminating these two populations (Figure 12). The properties which exceeded the 95% CI are:

MotifSampler difference score (property 4), distance to TSS (properties 10 and 11), repetitive element content (properties 12-15), minor and derived allele frequencies (properties 16 and 17), CpG content (property 19), DNaseI hypersensitive site content (property 20), PhastCons score (property 21) and phylogenetic tree distance (property 23). A concern with this analysis was that the background levels of properties assayed in individual promoter regions would not be comparable across promoters. To address this, a two-sampled t-test comparing the GROUP set was designed to identify discriminating properties within a gene (Figure 13). A complementary list of discriminating properties was identified which exceed the 95% CI. These properties were: MotifSampler difference scores (property 4), distance to TSS (properties 10 and 11), repetitive element content (properties 12-14), minor and derived allele frequency (properties 16 and 17), CpG element content (property 19), DNaseI hypersensitive site content (property 20), PhastCons conservation score (property 21), and phylogenetic tree distance (property 23).

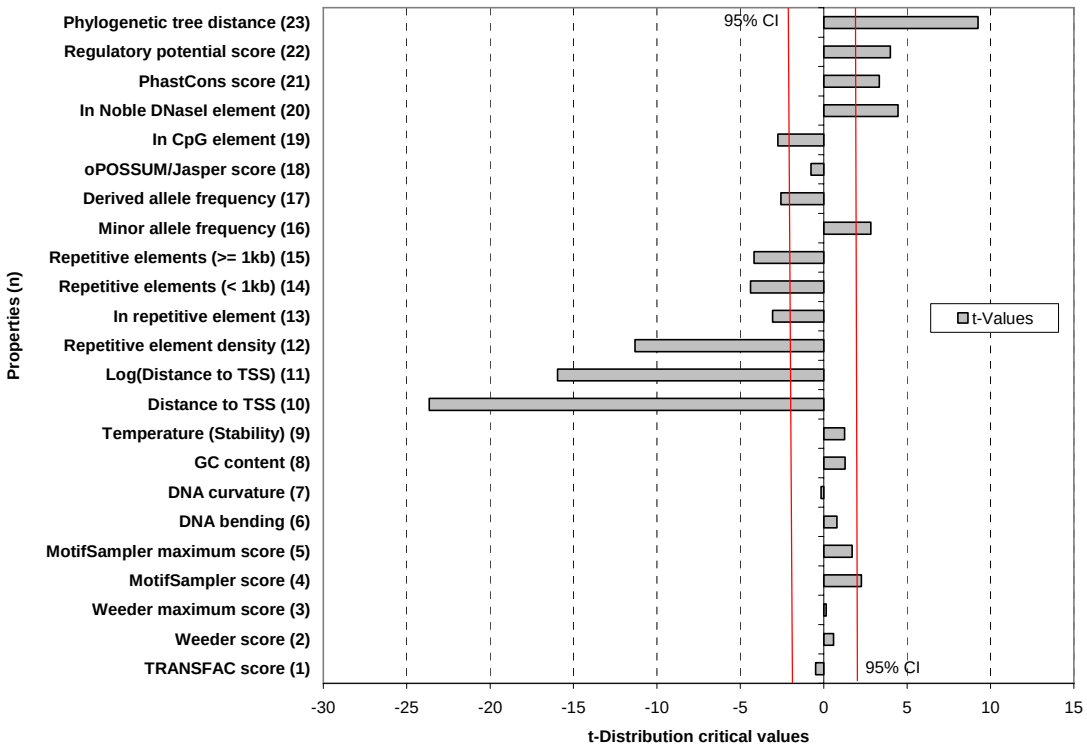

Figure 12: ALL analysis of 109-rSNPs against ufSNPs.

Individual results of each SNP from the 109-rSNP and ufSNP sets were analyzed using a two-sample t-test. A 95% CI was plotted for the minimum degrees of freedom from the properties (derived allele frequency; df=1227). All values extending beyond this threshold (to the left or right) have been identified as significant discriminants in this study. Properties that have negative t-distribution critical values indicate that 109-rSNP set has lower values than the ufSNP set (i.e. in property 10, distance to TSS, the negative value indicates lower distance values for the 109-rSNP set compared to the ufSNP set).

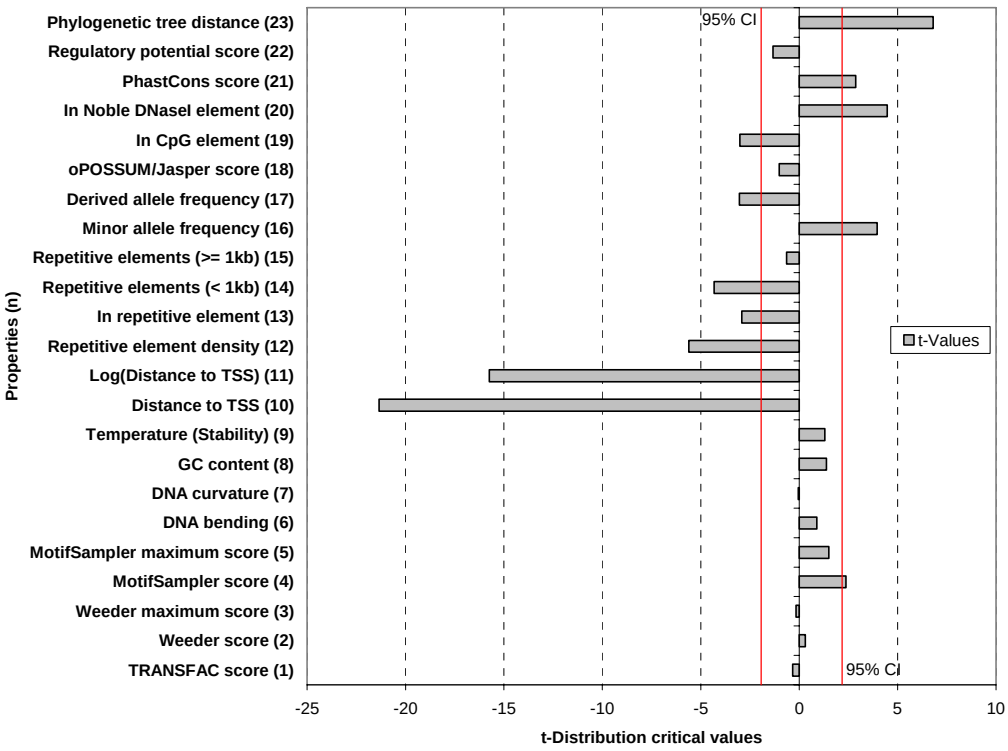

Figure 13: GROUP analysis of 109-rSNPs against ufSNPs.

Individual results of each SNP from the 109-rSNP and ufSNP sets were analyzed using a two-sample t-test. A 95% CI was plotted for the minimum degrees of freedom from the properties (derived allele frequency; df=1227). All values extending beyond this threshold (to the left or right) have been identified as significant discriminants in this study. Properties that have negative t-distribution critical values indicate that 109-rSNP set has negative deviation from average value for respective genes compared to the ufSNP set. (i.e. in property 10, distance to TSS, the negative value indicates lower distance values for the 109-rSNP set compared to the ufSNP set).

5.3.2 SVM cross-validation

We tested the classification performance of SVMs trained with the ALL and GROUP datasets via 10-fold cross-validation. For each SVM, the mean area under the ROC curve was 0.89 +/- 0.06 and 0.78 +/- 0.07, respectively. Both suggest good performance (Figure 14). It is notable that when removing distance from the classification the performance drops by only 5%.

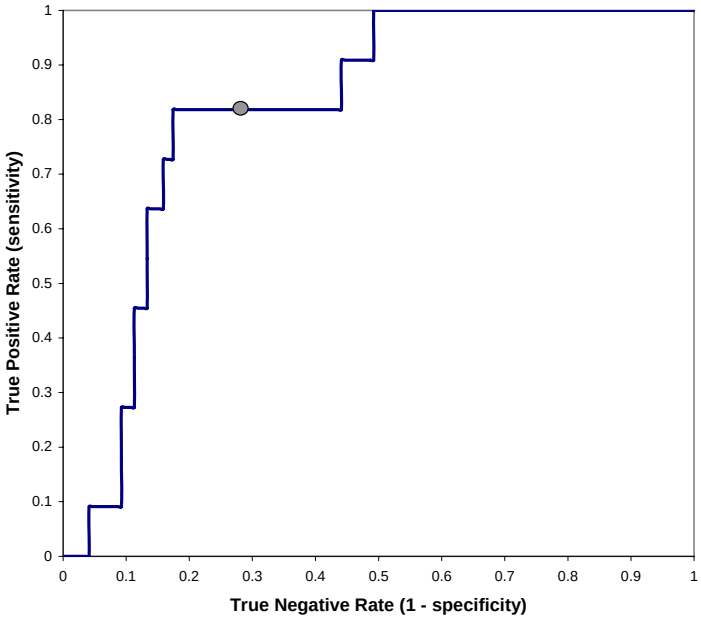

Figure 14: Receiver operating characteristic (ROC) curve for discriminating known regulatory SNPs from polymorphisms of unknown function.

This ROC curve was calculated by training an SVM on a randomly selected 90% subset of the 109-rSNP and ufSNP datasets. Here, 98 rSNPs and 2494 ufSNPs were utilized for training, followed by testing on the held-out 10%. The GROUP SVM approach was used for training. The area under this curve is 0.8196 which indicates very good performance. The dot marks the location of the decision boundary selected by the SVM. At this boundary, the SVM identifies 9 of 11 true positives and 146 of 195 true negatives.

5.3.3 Distance analysis

We investigate the effects of choosing a 6kb window in this study by identifying a window size that had equivalent average distances to TSS between the rSNP and ufSNP populations. A window size of 152bp satisfied this condition. At this window size only 16 rSNPs and 21 ufSNPs were available for analysis. When compared using the GROUP analysis t-test, only three properties were above the 95%CI: repetitive element density

(property 12), repeat content less than 1kb (property 14) and phylogenetic tree distance (property 23) (Figure 15).

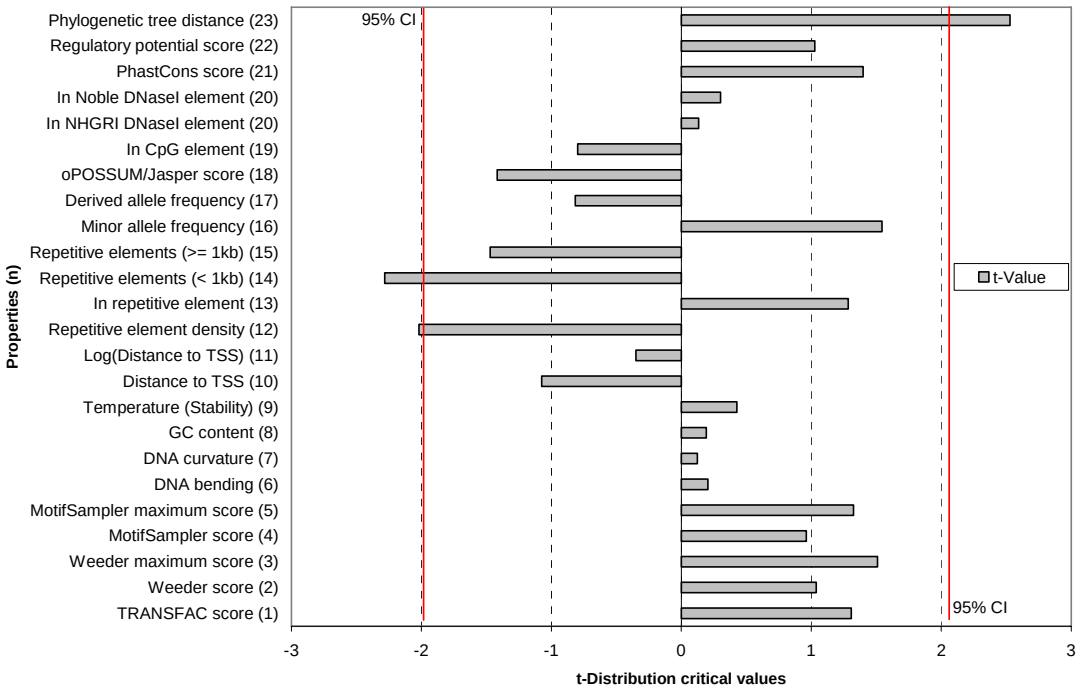

Figure 15: Equivalent distance distribution GROUP analysis of 16 rSNPs against 21 ufSNPs.

Individual results of each SNP from 16 rSNPs and 21 ufSNP sets were analyzed using a two-sample t-test. A 95% CI was plotted for the minimum degrees of freedom from the properties (minor allele frequency; df=26). All values extending beyond this threshold (to the left or right) have been identified as significant discriminants in this study. Properties that have negative t-distribution critical values indicate that rSNP set has negative deviation from average value for respective genes compared to the ufSNP set. (i.e. in property 10, distance to TSS, the negative value indicates lower distance values for the 109-rSNP set compared to the ufSNP set). SNPs were selected here from a reduced window size of 152bp. At this window size the average distance to TSS for the 16 rSNPs is equivalent to that of the 21 ufSNPs.

We also investigate what sorts of bias exists in the position of identified rSNPs. Our expectation was that well-established transcription factor binding sites such as the TATA- and CCAAT-boxes may be overrepresented and contribute to lower distance values. A histogram of rSNPs for the first 300bp of sequence from the TSS shows an

expected blip around the 21-31 position where 7 rSNPs are located, twice as many as average. However, it is apparent that these types of binding sites are only overrepresented slightly when compared to the distribution of rSNPs at other positions (Figure 16).

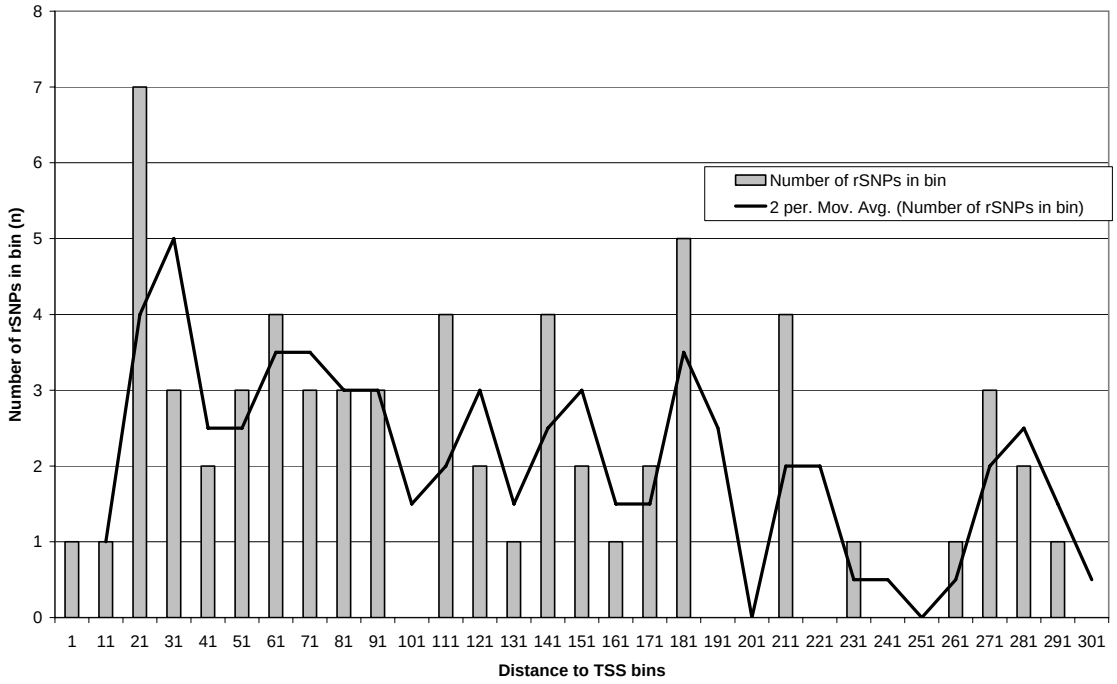

Figure 16: Histogram of positional bias of rSNPs for the first 300bp of sequence

The positions of rSNPs are plotted in a histogram for bin sizes of 10bp for the first 300bp of sequence from the transcription start site. A blip is seen at position 21-31 where it is likely that TATA and CCAAT-box binding sites are located. These types of rSNPs, however, are only slightly overrepresented in this study and from this graph it would not be expected to significantly bias the outcome.

5.3.4 Surveying new polymorphisms

We used this approach to prioritize SNPs in the proximal promoters of six different cancer-related genes. Two SVMs were constructed: the GROUP SVM and the GROUP SVM without the distance properties (properties 10 and 11). The latter was

constructed to mitigate the influence of distance to TSS on the classification. SNPs were manually identified that were high scoring using both SVMs. Of an additional 146 SNPs tested in 7 genes of importance to cancer and Huntington's disease, several features were observed: most predictions were proximal to the TSS but not in a ranked order as distance increases; some genes possessed very few positives while others were enriched with hits; and, most significantly, among the 146 SNPs tested, the 4th highest-scoring SNP has been recently shown at our institute to be in significant association with lymphoma (Brooks-Wilson, personal communication). This suggests that this method will be of great practical utility to prioritizing regulatory polymorphisms to assay for involvement in disease.

5.3.5 Availability

All pipeline software has been programmed in Perl and is available under the LGPL at http://www.bcgsc.ca under the name CHuM (Cis-acting Human mutation modules). All data is available from this site.

5.4 Discussion

Understanding the effects of sequence variation on gene expression remains a major challenge of modern genetics. The increasing practicality of genome-wide linkage and association studies coupled with new resources, such as the HapMap [295], offers significant potential to identify the allelic spectrum of common diseases. Furthermore, high-throughput categorization of allelic-specific expression differences through heterozygotes has suggested that somewhere between 25-50% of all human genes have

differential expression [369]. Once a regulatory haplotype is known, however, it is challenging to dissect the individual contributions of variants. To address this, as mentioned in the introduction to this chapter, computational analyses of non-coding variants have investigated distance to transcription start site, weight matrices and conservation-based approaches to prioritizing SNPs of unknown function. Recently, a study has investigated sequence composition around functional and non-functional SNPs for compositional differences indicative of their selective roles in transcription factor binding [390]. These authors reported 70% specificity and 20% sensitivity for 44 known regulatory polymorphisms when compared to non-functional polymorphisms. They note that using a broader range of properties, including distance to transcription start site and allele frequencies, may improve this performance.

This study introduces the largest publicly-available collection of regulatory polymorphisms--171 known regulatory polymorphisms from literature. Furthermore, this study investigates 109 regulatory polymorphisms (109-rSNP) and 2690 SNPs of unknown function (ufSNPs) in human core-promoter regions to identify properties that discriminate functional from non-functional polymorphisms.

Several properties were found which were above a 95% CI for discriminative potential.

Conservation-based metrics (Property 4, 21, 22 and 23)

Three different properties related to sequence conservation were identified as being significantly discriminatory in this study: MotifSampler difference score, PhastCons score and phylogenetic tree distance score. This result is not unexpected as

conservation is regarded as suggesting functional constraint. The latter two properties are not calculated in an allele-specific manner suggesting that regulatory polymorphisms are in intrinsically conserved regions. The MotifSampler and Weeder-based scores were calculated individually for each allele, to assess how changes in the patterns might affect their scoring. It is of note that Weeder, an enumerative approach, does not have significant allele-specific change in its scoring function while MotifSampler, an objective function algorithm, does; this result suggests that MotifSampler's optimization function is more sensitive to variance in motif composition. It is of further note that in the ALL analysis, the regulatory potential score has significance, yet in the the GROUP analysis, it does not. This suggests that there is less variance across the core-promoter for these types of scores. In total, these results underscore conservation as a useful discriminant of regulatory polymorphisms. However, it has previously been demonstrated that this conservation will not be sufficient to this purpose alone [136]--it is unlikely a regulatory region can be so conserved as to disallow polymorphism or so non-conserved as to have no functional constraint.

As an aside, the limits of conservation-based approaches were also illustrated by a supplementary activity of this thesis which was to investigate the evolution of the 109-rSNPs across several mammalian species. From this analysis we observed a rSNP in the gamma-globin gene which was in a primate-specific duplicated region; it is apparent that conservation scores based on alignments would not be able to detect conservation in this region while possibly some types of motif discovery algorithms should be able to identify the underlying motif (Figure 17). While only speculative at this point, one can

hypothesize that this duplication allowed relaxed selection on what appears to be a highly conserved motif.

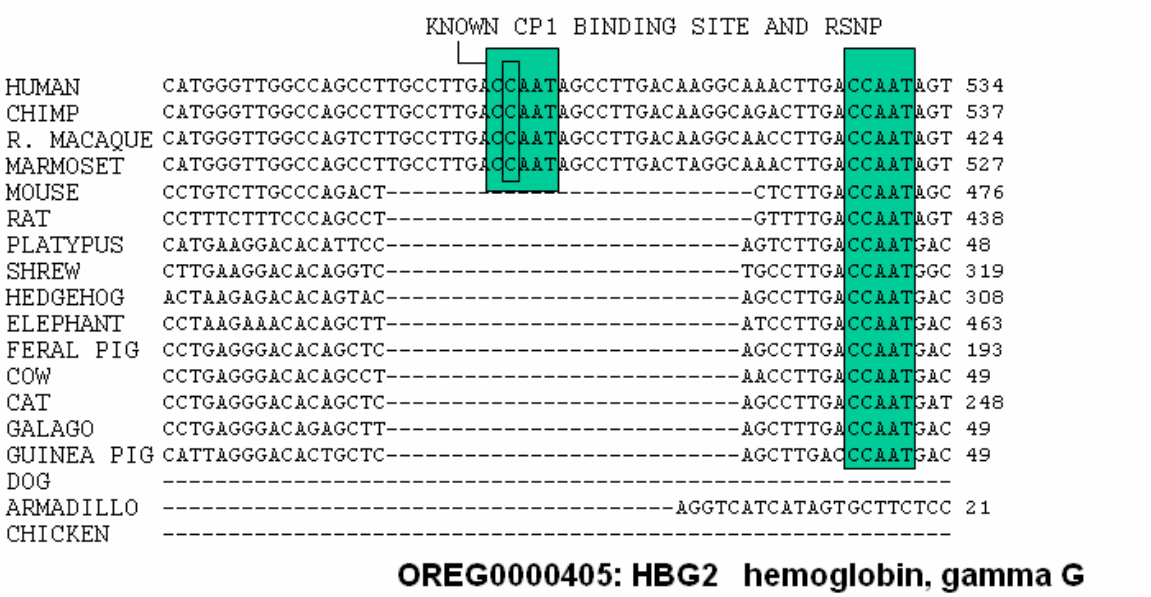

Figure 17: Mammalian alignment of OREG0000405

A ClustalW alignment of sequences around the OREG0000405 C/T human regulatory polymorphism in the HBG2 gene shows a primate specific duplication of a known CP1 transcription factor binding site. The duplicated motif is boxed and the regulatory polymorphism is in the sub-box. An alignment-based score around such a site would show little conservation whereas a motif discovery tool might pick up the essential CCAAT motif.

Distance to TSS (Property 10 and 11)

Distance to transcription start site has been identified as a successful discriminant in a previous study, where in 500bp assayed regions, 50% of the detected rSNPs were within 100bp of the TSS [509]. In this study, it is likely that ascertainment bias in the 109-rSNP set likely also contributes to the strength of this discriminant in this study.

Repetitive element content (Properties 12-15)

Repetitive content was investigated in this study because of its role in altering gene regulation and mirroring selective constraint in non-coding regions [255-258]. We

identified that regulatory SNPs were less likely to be in or around repetitive elements; not only was there less repetitive base content, but the number of distinct repetitive elements was also low. This suggests that regions that are under higher functional constraint are less likely to accrue repetitive elements and be subject to dysregulation. We also note that ascertainment bias by which 109-rSNPs were surveyed in terms of repetitive elements is not known. Of note, between the ALL and the GROUP analysis, the significance of the number of long repetitive elements collocated with a SNP is reduced. It is likely that these repeats extend over many of the SNPs in a group and, therefore, are averaged towards the null hypothesis.

Minor and derived allele frequency (Properties 16 and 17)

In this study, both minor and derived allele frequency were identified as significant discriminants. An interesting result though was that for genotyped SNPs, the minor allele frequency was higher in the 109-rSNP set than in the ufSNP set. Previous analyses of minor allele frequency have suggested that most functional SNPs are positioned around 6% [391] or possess no allele frequency bias [509]. In this study, the average minor allele frequency was approximately 22%. Since a subset of the 109-rSNP set has been derived from association studies, it is possible that ascertainment bias may explain part of this result as researchers may preferentially be choosing higher MAF SNPs because of their greater statistical power. Of interest, however, the derived allele frequency was lower in the 109-rSNP set than in the ufSNP set. This could suggest that many of the chimpanzee alleles have been driven to lower frequencies due to new variants increasing in frequency in our population either through population bottlenecks

or positive selection. The latter being a model of evolution of genetic susceptibility to common diseases explained by ancient alleles recently becoming disease predisposing due to changes in human lifestyle and life expectancy [515].

CpG elements (Properties 19)

Another interesting result was that SNPs in the 109-rSNP set were less likely to be in CpG islands than ufSNPs. It is likely that an admixture of CpG islands and CpG-less, TATA promoters caused this result; both types of promoter are known to be prevalent in the genome [40,43]. Since, CpG expectancy was normalized from average values at specific distances from the TSS of associated genes across individual chromosomes, an admixture of a CpG and CpG-less promoters would drive the 109-rSNP values lower than the ufSNP values (Figure 12).

DNaseI hypersensitive sites (Properties 20)

In this study, the 109-rSNPs were more likely to be in predicted hypersensitive sites than the ufSNPs. This result was expected since DNaseI hypersensitivity is well-established as being associated with DNA-protein binding.

Several properties were not above our significance threshold in this study. Of interest, both weight matrix-based approaches did not discriminate well. Additionally, our definition of coexpression was significantly broad as to allow multiple coexpressed partners for any given gene; this may have reduced the overall effectiveness of reducing transcription factor binding profiles using this information. However, the performance of

the coexpression filtered approach using oPOSSUM was moderately better than the TRANSFAC approach alone. This suggests that targetted analysis of specific, biologically-relevant transcription factors may further increase the discriminating ability of this approach. This should also act as a warning to those that have in the past applied the TRANSFAC approach to this problem indiscriminately. Furthermore, none of the DNA structural or stability analyses were successfully discriminatory. This analysis can be interpreted as, not only do these features have non-generalizable allele-specific effects using the datasets in this study, but since these analyses measure local sequence composition, nothing is particularly important in terms of specific base changes.

A principal goal of this study was to be able to use the 109-rSNP and ufSNP set to train two SVMS to discriminate polymorphisms in novel genes. It is of note, despite certain ascertainment biases, that our sensitivity and specificity for the ALL test is 0.85 +/- 0.12 and 0.80 +/- 0.06, respectively, and for GROUP test is 0.74 +/- 0.19 and 0.65 +/- 0.14, respectively. This performance only drops by 5% when distance to TSS is removed from the analysis. While steps were taking to explicity control for distance bias, some properties may be indirectly correlated and future analyses with enriched datasets of regulatory polymorphisms may be able to dissect these interactions. Of fundamental necessity to these future analyses is a dataset of core-promoter polymorphisms that are non-functional across a broad range of cell types; since our negative control set was a neutral set, it is assured that more accurate performance metrics can come from addition of a reliable negative control set. Despite this fact, using a neutral set, we can suggest that to achieve perfect performance approximately 1/5 of the ufSNPs in core promoter regions across the genome must be regulatory polymorphisms (that is, if at the SVMs

decision plane, all the false positives are actually true positives). While certainly speculative, this seems not unreasonable as recent dissection of a complex haplotype has demonstrated that some traits are the product of the effects of several regulatory SNPs (the authors reported that 5 of 29 SNPs tested contributed to the variation in expression) [396].

In summary, this study introduces a new dataset of for the investigation of regulatory polymorphisms. These results describe the utility of different gene regulation and population genetics properties in classifying regulatory polymorphisms. Finally, we have also introduced the first gene regulation and population genetics-based approach to classifying these polymorphisms in the core promoters of human genes.

5.5 Conclusions

This work introduces a novel dataset and approach for categorizing regulatory polymorphisms using the ORegAnno technology developed in Chapter 4. Several future directions are likely to be investigated as part of this work. Specifically, several SNPs of disease relevance are being experimentally tested for function to independently validate this approach. Additionally, a logical extension of this work is to begin to categorize regulatory polymorphisms outside of the core promoter or within haplotypes. Our ability to do this, however, is challenged by the lack of SNPs characterizing regulatory elements outside of these regions and the length of linkage disequilibrium in some regions. The development of the ORegAnno resources will have future significance in assembling these datasets and will further allow direct investigation of polymorphisms in known

regulatory elements. The next and final chapter summarizes the major conclusions of this thesis.

Chapter 6: Summary and Conclusions

6.1 Summary

Identification of the mechanisms by which genes are regulated in eukaryotes is one of the principal challenges of modern biology. New insight into these mechanisms and their locations of action will enable researchers to answer fundamental questions related to evolution and development. Understanding the role that heritable mutation plays on the mechanisms of gene regulation will further enable future clinicians to identify and potentially treat various human diseases and improve patient quality of life as such mutations have been identified in the etiology of diseases like cancer [397,398], depression [399], systematic lupus erythematosus [400], perinatal HIV-1 transmission [401], and response to type 1 interferons [402].

This thesis describes novel strategies for computational analysis of regulatory elements and regulatory polymorphisms. Additionally, this work generates novel hypotheses regarding the use and assessment of bioinformatics data and tools. A major biological hypothesis of this work was that regulatory polymorphisms could be predicted computationally using diverse genome annotation, *in silico* gene regulatory analyses and population genetics parameters. The aim of this work was to develop an approach which allowed investigation of this hypothesis and was amenable to making novel predictions of new regulatory polymorphisms. A major challenge in achieving this aim has been the lack of known transcription factor binding sites and regulatory polymorphisms. Furthermore, at the onset of this thesis, very little was known in literature about the relative merits of particular regulatory element tools, despite a burgeoning population of such tools. Both difficulties were addressed in my studies. A major biological result of this work has been a characterization of discriminating properties in a novel dataset of

regulatory polymorphisms hand-curated from literature. Additionally, throughout this thesis, this work has had a specific focus on addressing biological challenges by using new technology that may have a lasting significance on the wider research community. The results of this focus have been novel research into the application of 3D technology to genome visualization, peer-to-peer technology for resource discovery and comparison, and a novel community-based system for regulatory element annotation.

6.2 Predicting regulatory elements

For accurate prediction of regulatory elements, new strategies are required that integrate diverse sources of annotation and make use of enriched biological information, such as coexpression and genome conservation data. Also new approaches are required for measuring the performance of these strategies. These principles are well-established in modern genomics as major international collaboration under the ENCODE project is dedicated to developing these strategies. In this thesis, I have addressed some of the major challenges with modern regulatory element prediction with the development of novel resources.

The Sockeye tool (Chapter 2) was developed with a principal aim of being able to compare sequence and annotation across orthologous gene sets (and then coexpressed gene sets). This tool was integrated with the ability to perform regulatory element analyses using weight matrix-based and *de novo* motif discovery approaches on demand. These features made it the first tool to provide on demand regulatory element analyses in the context of rich genome annotation. As part of this visualization, we used 3D technology to exploit the extra dimension's ability to compact information and represent

scores as height; the height dimension allowed for score-based identification of CACNL1AS/Cacna1s muscle and liver-specific regulatory modules in the context of their known gene annotation. The parallel visualization of multiple genomes in 3D had the further benefit of allowing the comparative visualization of coronavirus genomes from which a qualitative assessment of the evolutionary history of the SARs coronavirus could be made. While 3D usage was a serious design decision with practical advantage in performing comparative analyses, it was a major paradigm shift at the time as most data was (and remains) solely displayed in static web-pages. Since this research was conducted, adoption of this technology for comparative genomics has not been widespread; it is likely that ease of data access (as Sockeye is a standalone application) and human factors associated with adapting to a 3D environment impede its adoption.

During the development of Sockeye, I became concerned with existing strategies for integrating bioinformatics resources and the practical assessment of a wide-range of bioinformatics methodology. The Chinook tool (Chapter 3) was developed using decentralized peer-to-peer technology to address several general challenges that exist in bioinformatics, namely the disunified distribution and usage of bioinformatics resources, and, specifically, to create a benchmark by which diverse bioinformatics utilities could be evaluated. Chinook presented a novel strategy for bioinformatics research--decentralized peer-to-peer technology could freely propagate the existence of and facilitate the usage and comparison of new and well-known bioinformatics resources to a broad spectrum of researchers without requiring centralized authority or resources. This need has been reiterated in various forms in high-profile commentaries [494,516]. Despite the potential of this technology and that of complementary integrative methodologies like BioMOBY

[517], there is no major impetus to achieve a higher-level of resource interoperability at this time making such resources of utility to interested bioinformatics sub-communities only. For this thesis, however, using Chinook, I was further able to integrate a dozen different regulatory analysis tools and create a benchmark web application for analyzing their performance called the Open Regulatory Competition (ORC). ORC was designed to provide statistics of comparison (sensitivity, specificity, positive predictive value and correlation coefficients) found in a recently published study by Tompa et al [97]. ORC is currently limited as it only allows individual genomic regions to be utilized as positive controls. Its development was primarily a proof of principle as to how open benchmarks can be created for sustained analysis of these types of tools. It is planned that future enhancement of ORC using ORegAnno datasets will act as a catalyst to wider integration of motif discovery tools and a more recognizable benchmark for assessing their performance.

Through developing ORC, I became interested in aggregating known regulatory elements and polymorphisms to, not only assess the performance of motif discovery tools, but to measure their performance with introduced regulatory polymorphisms. However, the availability of curated regulatory elements and regulatory polymorphisms at the time was markedly poor. To address this, I designed a unique resource for community annotation called ORegAnno (Chapter 4) which would serve several purposes. First, it would allow researchers to deposit their collections of regulatory elements. Second, it would introduce new mechanisms for maintaining the validity of such annotations. Third, the data could be extracted in a computer amenable format. This project has addressed a long-desired need in the genomics community and since its

development it has achieve considerable interest and support. At this point in time, several new datasets are being added to ORegAnno and an international collaboration has been developed with the specific aim of populating this resource (a jamboree called RegCreative will be held in the winter of 2006). It is my future plan to be able to use this to investigate the distribution of regulatory polymorphisms in experimentally-validated transcription factor binding sites.

A major component of this thesis was introducing new paradigms for usage, comparison and visualization of gene regulatory data and computional analysis tools. Modern genetics analysis has an increasing dependence on the efficacy of bioinformatics approaches. Recognizing the value of such research will enhance future biological hypotheses and inferences.

6.3 Prediction regulatory polymorphisms

The principal biological aim throughout the course of this thesis was to investigate strategies for predicting regulatory polymorphisms. At the onset of this thesis, there existed no bioinformatics resources or generalized methodology amenable to this analysis. One of my principle focuses in developing Sockeye (Chapter 1) was to integrate regulatory prediction techniques with genetic variant information; in this regard, Sockeye became the first tool that allowed the seamless computional analysis of motifs in the context of genetic variant information. During the course of this thesis, studies started to become available introducing collections of regulatory polymorphisms and investigations of particular features of these SNPs. Of note, were comprehensive surveys by Rockman and Wray [377], and the characterization of several new rSNPs by Buckland

et al. [509]. The increasing availability of such annotation made it feasible to consider designing a computational approach to collating and then analyzing the properties of these SNPs. This was also further rationale behind the design of ORegAnno (Chapter 4) as it would allow the computer-amenable curation of rSNPs. To populate the ORegAnno resource, I conducted an extensive survey of the literature to curate a collection of known regulatory polymorphisms. The ORegAnno collection now has 174 regulatory polymorphisms with known effects on gene expression--the largest openly-access, computationally-amenable collection to date. Using the ORegAnno rSNP collection, I investigated a broad range of properties that may discriminate them from the null hypothesis in the core promoter of associated human genes (Chapter 5). 109 rSNPs and 2690 SNPs of unknown function were categorized and several properties were found of significance (above a 95% CI) in discriminating these populations. Among those properties were: MotifSampler difference score, distance to transcription start site, repetitive element content, minor and derived allele frequencies, CpG content, PhastCons score and phylogenetic tree distance (property 23)—the most novel of which is the repetitive element content properties, which suggest that regulatory polymorphisms are less likely to occur in areas where many small repeats have accrued. Using these discriminants, I was able to train an SVM to predict with 85% sensitivity and 80% specificity known regulatory polymorphisms within this set. This approach is the first multiple-parameter approach to regulatory polymorphism analysis. Furthermore, this approach has been applied to discovering new functional regulatory polymorphisms in several cancer genes; an investigation of non-coding SNPs in one such gene has identified one high-scoring rSNP that has also been recognized as being in association

with lymphoma through experiments in our lab (Brooks-Wilson, personal communication). Further investigations are being conducted of other high-scoring SNPs.

6.4 Conclusions

Bioinformatics approaches offer the potential to discover regulatory elements and the effects of genetic variation on them. These analyses unequivocally are more reliable when performed using well-guided biological inferences. Analyses of computational tools and data designed for this purpose has identified several weaknesses, namely the lack of benchmarks for managing an extensive set of available regulatory tools and well-documented and computationally-amenable gene regulation databases. Community-based strategies offer enriched opportunity to guide these types of analyses in the future. Additionally, the construction of such resources has had practical utility for investigating the properties of regulatory polymorphisms. The development of the ORegAnno resource has facilitated the development of an approach that allows for discrimination of regulatory polymorphisms in the core promoter region of human genes using regulatory and population genetic parameters. Computational analysis of regulatory polymorphisms still remains a relatively unexplored field of inquiry. It is inevitable as modern genetics moves towards whole-genome association and linkage assays, the enhancement of such strategies will have fundamental importance to uncovering the etiology of various determinants of health.

References

1. Mendel G (1865) Versuche über Pflanzen-Hybriden. Verh naturf Ver Brunn 4: 3-47.
2. Avery OT, MacLeod CM, McCarty M (1944) Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III. Journal of Experimental Medicine 79: 137-158.
3. Hershey AD, Chase M (1952) Independent functions of viral protein and nucleic acid in growth of bacteriophage. J Gen Physiol 36: 39-56.
4. Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171: 737-738.
5. Maddox B (2003) The double helix and the 'wronged heroine'. Nature 421: 407-408.
6. Morgan TH (1910) Sex Limited Inheritance in Drosophila. Science 32: 120-122.
7. Beadle GW, Tatum EL (1941) Genetic control of biochemical reactions in Neurospora. PNAS 27: 499-506.
8. Nirenberg M (2004) Historical review: Deciphering the genetic code--a personal account. Trends Biochem Sci 29: 46-54.
9. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, et al. (1977) Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265: 687-695.
10. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.
11. Chan EY (2005) Advances in sequencing technology. Mutat Res 573: 13-40.
12. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931-945.
13. Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. Nucleic Acids Res 34: D332-334.
14. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520-562.
15. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428: 493-521.
16. The Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69-87.
17. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438: 803-819.
18. Schwarz EM, Antoshechkin I, Bastiani C, Bieri T, Blasiar D, et al. (2006) WormBase: better software, richer content. Nucleic Acids Res 34: D475-478.
19. Grumbling G, Strelets V (2006) FlyBase: anatomical data, images and queries. Nucleic Acids Res 34: D484-488.
20. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, et al. (2006) The UCSC Genome Browser Database: update 2006. Nucleic Acids Res 34: D590-598.

21. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, et al. (2006) Ensembl 2006. Nucleic Acids Res 34: D556-561.
22. Furey TS (2006) Comparison of human (and other) genome browsers. Hum Genomics 2: 266-270.
23. Teufel A, Krupp M, Weinmann A, Galle PR (2006) Current bioinformatics tools in genomic biomedical research (Review). Int J Mol Med 17: 967-973.
24. Webb S (1976) Nutrition, time and motion in metabolism and genetics. Charles C Thomas Publisher.
25. Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3: 318-356.
26. Monod J (1942) Recherches sur la croissance des cellules bactériennes. PhD thesis Actualités scientifiques et industrielles, Hermann, Paris.
27. Lewis M (2005) The lac repressor. C R Biol 328: 521-548.
28. Ross J (1996) Control of messenger RNA stability in higher eukaryotes. Trends Genet 12: 171-175.
29. Almeida R, Allshire RC (2005) RNA silencing and genome regulation. Trends Cell Biol 15: 251-258.
30. Robertson KD (2005) DNA methylation and human disease. Nat Rev Genet 6: 597-610.
31. Chen ZX, Riggs AD (2005) Maintenance and regulation of DNA methylation patterns in mammals. Biochem Cell Biol 83: 438-448.
32. Holmes R, Soloway PD (2006) Regulation of imprinted DNA methylation. Cytogenet Genome Res 113: 122-129.
33. Attwood JT, Yung RL, Richardson BC (2002) DNA methylation and the regulation of gene transcription. Cell Mol Life Sci 59: 241-257.
34. Mizzen CA, Allis CD (1998) Linking histone acetylation to transcriptional regulation. Cell Mol Life Sci 54: 6-20.
35. Bashirullah A, Cooperstock RL, Lipshitz HD (2001) Spatial and temporal control of RNA stability. Proc Natl Acad Sci U S A 98: 7025-7028.
36. Capelson M, Corces VG (2004) Boundary elements and nuclear organization. Biol Cell 96: 617-629.
37. Udvardy A (1999) Dividing the empire: boundary chromatin elements delimit the territory of enhancers. Embo J 18: 1-8.
38. Arnone MI, Davidson EH (1997) The hardwiring of development: organization and function of genomic regulatory systems. Development 124: 1851-1864.
39. Landry JR, Mager DL, Wilhelm BT (2003) Complex controls: the role of alternative promoters in mammalian genomes. Trends Genet 19: 640-648.
40. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, et al. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38: 626-635.
41. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. (2005) The transcriptional landscape of the mammalian genome. Science 309: 1559-1563.
42. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, et al. (2005) Antisense transcription in the mammalian transcriptome. Science 309: 1564-1566.

43. Saxonov S, Berg P, Brutlag DL (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A 103: 1412-1417.
44. Gross P, Oelgeschlager T (2006) Core promoter-selective RNA polymerase II transcription. Biochem Soc Symp: 225-236.
45. Muller F, Tora L (2004) The multicoloured world of promoter recognition complexes. Embo J 23: 2-8.
46. Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 8: 1499-1504.
47. Antequera F, Bird A (1999) CpG islands as genomic footprints of promoters that are associated with replication origins. Curr Biol 9: R661-667.
48. Johnson KD, Christensen HM, Zhao B, Bresnick EH (2001) Distinct mechanisms control RNA polymerase II recruitment to a tissue-specific locus control region and a downstream promoter. Mol Cell 8: 465-471.
49. Louie MC, Yang HQ, Ma AH, Xu W, Zou JX, et al. (2003) Androgen-induced recruitment of RNA polymerase II to a nuclear receptor-p160 coactivator complex. Proc Natl Acad Sci U S A 100: 2226-2230.
50. Shang Y, Myers M, Brown M (2002) Formation of the androgen receptor transcription complex. Mol Cell 9: 601-610.
51. Spicuglia S, Kumar S, Yeh JH, Vachez E, Chasson L, et al. (2002) Promoter activation by enhancer-dependent and -independent loading of activator and coactivator complexes. Mol Cell 10: 1479-1487.
52. Szutorisz H, Dillon N, Tora L (2005) The role of enhancers as centres for general transcription factor recruitment. Trends Biochem Sci 30: 593-599.
53. Zhao H, Dean A (2005) Organizing the genome: enhancers and insulators. Biochem Cell Biol 83: 516-524.
54. Li Q, Peterson KR, Fang X, Stamatoyannopoulos G (2002) Locus control regions. Blood 100: 3077-3086.
55. Bellen HJ (1999) Ten years of enhancer detection: lessons from the fly. Plant Cell 11: 2271-2281.
56. Andrioli LP, Vasisht V, Theodosopoulou E, Oberstein A, Small S (2002) Anterior repression of a Drosophila stripe enhancer requires three position-specific mechanisms. Development 129: 4931-4940.
57. Messina DN, Glasscock J, Gish W, Lovett M (2004) An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. Genome Res 14: 2041-2047.
58. The ENCODE Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306: 636-640.
59. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM (2006) Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. Genome Res 16: 1-10.
60. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, et al. (2000) Genome-wide location and function of DNA binding proteins. Science 290: 2306-2309.
61. Bieda M, Xu X, Singer MA, Green R, Farnham PJ (2006) Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. Genome Res 16: 595-605.

62. Carroll JS, Liu XS, Brodsky AS, Li W, Meyer CA, et al. (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. Cell 122: 33-43.
63. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, et al. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell 116: 499-509.
64. Euskirchen G, Royce TE, Bertone P, Martone R, Rinn JL, et al. (2004) CREB binds to multiple loci on human chromosome 22. Mol Cell Biol 24: 3804-3814.
65. Scacheri PC, Davis S, Odom DT, Crawford GE, Perkins S, et al. (2006) Genome-wide analysis of menin binding provides insights into MEN1 tumorigenesis. PLoS Genet 2: e51.
66. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, et al. (2005) A high-resolution map of active promoters in the human genome. Nature 436: 876-880.
67. Beyer A, Workman C, Hollunder J, Radke D, Moller U, et al. (2006) Integrated assessment and prediction of transcription factor binding. PLoS Comput Biol 2: e70.
68. Euskirchen G, Snyder M (2004) A plethora of sites. Nat Genet 36: 325-326.
69. Wei CL, Wu Q, Vega VB, Chiu KP, Ng P, et al. (2006) A global map of p53 transcription-factor binding sites in the human genome. Cell 124: 207-219.
70. Sabo PJ, Humbert R, Hawrylycz M, Wallace JC, Dorschner MO, et al. (2004) Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. Proc Natl Acad Sci U S A 101: 4537-4542.
71. Crawford GE, Holt IE, Mullikin JC, Tai D, Blakesley R, et al. (2004) Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. Proc Natl Acad Sci U S A 101: 992-997.
72. Dorschner MO, Hawrylycz M, Humbert R, Wallace JC, Shafer A, et al. (2004) High-throughput localization of functional elements by quantitative chromatin profiling. Nat Methods 1: 219-225.
73. Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5: 276-287.
74. Siggia ED (2005) Computational methods for transcriptional regulation. Curr Opin Genet Dev 15: 214-221.
75. Nardone J, Lee DU, Ansel KM, Rao A (2004) Bioinformatics for the 'bench biologist': how to find regulatory regions in genomic DNA. Nat Immunol 5: 768-774.
76. Rombauts S, Florquin K, Lescot M, Marchal K, Rouze P, et al. (2003) Computational approaches to identify promoters and cis-regulatory elements in plant genomes. Plant Physiol 132: 1162-1176.
77. MacIsaac KD, Fraenkel E (2006) Practical strategies for discovering regulatory DNA sequence motifs. PLoS Comput Biol 2: e36.
78. Pedersen AG, Baldi P, Chauvin Y, Brunak S (1999) The biology of eukaryotic promoter prediction--a review. Comput Chem 23: 191-207.
79. Ohler U, Niemann H (2001) Identification and analysis of eukaryotic promoters: recent computational approaches. Trends Genet 17: 56-60.
80. Pavesi G, Mauri G, Pesole G (2004) In silico representation and discovery of transcription factor binding sites. Brief Bioinform 5: 217-236.

81. Sandve GK, Drablos F (2006) A survey of motif discovery methods in an integrated framework. Biol Direct 1: 11.
82. Bajic VB, Tan SL, Christoffels A, Schonbach C, Lipovich L, et al. (2006) Mice and men: their promoter properties. PLoS Genet 2: e54.
83. Costantini M, Clay O, Auletta F, Bernardi G (2006) An isochore map of human chromosomes. Genome Res 16: 536-541.
84. Vinogradov AE (2005) Dualism of gene GC content and CpG pattern in regard to expression in the human genome: magnitude versus breadth. Trends Genet 21: 639-643.
85. Vinogradov AE (2005) Noncoding DNA, isochores and gene expression: nucleosome formation potential. Nucleic Acids Res 33: 559-563.
86. Antequera F (2003) Structure, function and evolution of CpG island promoters. Cell Mol Life Sci 60: 1647-1658.
87. Bird A (1999) DNA methylation de novo. Science 286: 2287-2288.
88. Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. J Mol Biol 196: 261-282.
89. Pfeifer GP (2006) Mutagenesis at methylated CpG sequences. Curr Top Microbiol Immunol 301: 259-281.
90. Aerts S, Thijs G, Dabrowski M, Moreau Y, De Moor B (2004) Comprehensive analysis of the base composition around the transcription start site in Metazoa. BMC Genomics 5: 34.
91. Down TA, Hubbard TJ (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res 12: 458-461.
92. Ponger L, Mouchiroud D (2002) CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. Bioinformatics 18: 631-633.
93. Davuluri RV, Grosse I, Zhang MQ (2001) Computational identification of promoters and first exons in the human genome. Nat Genet 29: 412-417.
94. Ioshikhes IP, Zhang MQ (2000) Large-scale human promoter mapping using CpG islands. Nat Genet 26: 61-63.
95. Goldberg ML (1979) Ph.D. Thesis. Stanford University.
96. Lawrence CE, Reilly AA (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins 7: 41-51.
97. Tompa M, Li N, Bailey TL, Church GM, De Moor B, et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23: 137-144.
98. Rigoutsos I, Floratos A (1998) Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. Bioinformatics 14: 55-67.
99. Frith MC, Hansen U, Spouge JL, Weng Z (2004) Finding functional sequence elements by multiple local alignment. Nucleic Acids Res 32: 189-200.
100. Workman CT, Stormo GD (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. Pac Symp Biocomput: 467-478.

101. Pavesi G, Mereghetti P, Mauri G, Pesole G (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res 32: W199-203.
102. Hughes JD, Estep PW, Tavazoie S, Church GM (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol 296: 1205-1214.
103. Liu XS, Brutlag DL, Liu JS (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat Biotechnol 20: 835-839.
104. Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, et al. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics 17: 1113-1122.
105. Bailey T, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, California: 28-36.
106. Hertz GZ, Hartzell GW, 3rd, Stormo GD (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. Comput Appl Biosci 6: 81-92.
107. ELPH (2006) http://www.cbcb.umd.edu/software/ELPH/.
108. Thompson W, Rouchka EC, Lawrence CE (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. Nucleic Acids Res 31: 3580-3585.
109. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, et al. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262: 208-214.
110. Palaniswamy SK, James S, Sun H, Lamb RS, Davuluri RV, et al. (2006) AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks. Plant Physiol 140: 818-829.
111. AtProbe (2006) http://rulai.cshl.edu/software/index1.htm.
112. CEPDB (2006) http://rulai.cshl.edu/software/index1.htm.
113. Kummerfeld SK, Teichmann SA (2006) DBD: a transcription factor prediction database. Nucleic Acids Res 34: D74-81.
114. Bergman CM, Carlson JW, Celniker SE (2005) Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. Bioinformatics 21: 1747-1749.
115. Schmid CD, Perier R, Praz V, Bucher P (2006) EPD in its twentieth year: towards complete promoter coverage of selected model organisms. Nucleic Acids Res 34: D82-85.
116. Adryan B, Teichmann SA (2006) FlyTF: a systematic review of site-specific transcription factors in the fruit fly Drosophila melanogaster. Bioinformatics 22: 1532-1533.
117. Pohar TT, Sun H, Davuluri RV (2004) HemoPDB: Hematopoiesis Promoter Database, an information resource of transcriptional regulation in blood cell development. Nucleic Acids Res 32: D86-90.

118. Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, et al. (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. Nucleic Acids Res 34: D95-97.
119. LSPD (2006) http://rulai.cshl.edu/software/index1.htm.
120. Zhang C, Xuan Z, Otto S, Hover JR, McCorkle SR, et al. (2006) A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. Nucleic Acids Res 34: 2238-2246.
121. Sun H, Palaniswamy SK, Pohar TT, Jin VX, Huang TH, et al. (2006) MPromDb: an integrated resource for annotation and visualization of mammalian gene promoters and ChIP-chip experimental data. Nucleic Acids Res 34: D98-103.
122. Palaniswamy SK, Jin VX, Sun H, Davuluri RV (2005) OMGProm: a database of orthologous mammalian gene promoters. Bioinformatics 21: 835-836.
123. Grienberg I, Benayahu D (2005) Osteo-Promoter Database (OPD) -- promoter analysis in skeletal cells. BMC Genomics 6: 46.
124. Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res 27: 297-300.
125. Lescot M, Dehais P, Thijs G, Marchal K, Moreau Y, et al. (2002) PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res 30: 325-327.
126. Shahmuradov IA, Gammerman AJ, Hancock JM, Bramley PM, Solovyev VV (2003) PlantProm: a database of plant promoter sequences. Nucleic Acids Res 31: 114-117.
127. Zhu J, Zhang MQ (1999) SCPD: a promoter database of the yeast Saccharomyces cerevisiae. Bioinformatics 15: 607-611.
128. Gallo SM, Li L, Hu Z, Halfon MS (2006) REDfly: a Regulatory Element Database for Drosophila. Bioinformatics 22: 381-383.
129. Kanamori M, Konno H, Osato N, Kawai J, Hayashizaki Y, et al. (2004) A genome-wide and nonredundant mouse transcription factor database. Biochem Biophys Res Commun 322: 787-793.
130. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, et al. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34: D108-110.
131. Zhao F, Xuan Z, Liu L, Zhang MQ (2005) TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies. Nucleic Acids Res 33: D103-107.
132. Kolchanov NA, Ignatieva EV, Ananko EA, Podkolodnaya OA, Stepanenko IL, et al. (2002) Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucleic Acids Res 30: 312-317.
133. Zhang MQ (1998) A discrimination study of human core-promoters. Pac Symp Biocomput: 240-251.
134. Xuan Z, Zhao F, Wang J, Chen G, Zhang MQ (2005) Genome-wide promoter extraction and analysis in human, mouse, and rat. Genome Biol 6: R72.
135. Gershenzon NI, Ioshikhes IP (2005) Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. Bioinformatics 21: 1295-1300.
136. Hoogendoorn B, Coleman SL, Guy CA, Smith K, Bowen T, et al. (2003) Functional analysis of human promoter polymorphisms. Hum Mol Genet 12: 2249-2254.

137. MotifSampler (2006) http://homes.esat.kuleuven.be/~thijs/Work/MotifSampler.html.
138. Klug SJ, Famulok M (1994) All you wanted to know about SELEX. Mol Biol Rep 20: 97-107.
139. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A (1982) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Res 10: 2997-3011.
140. Fickett JW (1996) Quantitative discrimination of MEF2 sites. Mol Cell Biol 16: 437-441.
141. D'Haeseleer P (2006) What are DNA sequence motifs? Nat Biotechnol 24: 423-425.
142. Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18: 6097-6100.
143. Stormo GD (1998) Information content and free energy in DNA--protein interactions. J Theor Biol 195: 135-137.
144. Zhou Q, Liu JS (2004) Modeling within-motif dependence for transcription factor binding site predictions. Bioinformatics 20: 909-916.
145. Hannenhalli S, Wang LS (2005) Enhanced position weight matrices using mixture models. Bioinformatics 21 Suppl 1: i204-212.
146. Lenhard B, Wasserman WW (2002) TFBS: Computational framework for transcription factor binding site analysis. Bioinformatics 18: 1135-1136.
147. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14: 1188-1190.
148. Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, et al. (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol 203: 439-455.
149. Pennacchio LA, Rubin EM (2001) Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet 2: 100-109.
150. Gumucio DL, Heilstedt-Williamson H, Gray TA, Tarle SA, Shelton DA, et al. (1992) Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. Mol Cell Biol 12: 4919-4929.
151. Duret L, Bucher P (1997) Searching for regulatory elements in human noncoding sequences. Curr Opin Struct Biol 7: 399-406.
152. Ureta-Vidal A, Ettwiller L, Birney E (2003) Comparative genomics: genome-wide analysis in metazoan eukaryotes. Nat Rev Genet 4: 251-262.
153. Solovyev VV, Shahmuradov IA (2003) PromH: Promoters identification using orthologous genomic sequences. Nucleic Acids Res 31: 3540-3545.
154. Zhang MQ (2003) Prediction, annotation, and analysis of human promoters. Cold Spring Harb Symp Quant Biol 68: 217-225.
155. Hardison RC (2000) Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet 16: 369-372.
156. Cooper GM, Sidow A (2003) Genomic regulatory regions: insights from comparative sequence analysis. Curr Opin Genet Dev 13: 604-610.
157. Nobrega MA, Pennacchio LA (2004) Comparative genomic analysis as a tool for biological discovery. J Physiol 554: 31-39.

158. Tompa M (2001) Identifying functional elements by comparative DNA sequence analysis. Genome Res 11: 1143-1144.
159. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, et al. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424: 788-793.
160. Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC (2003) Cross-species sequence comparisons: a review of methods and available resources. Genome Res 13: 1-12.
161. Margulies EH, Blanchette M, Haussler D, Green ED (2003) Identification and characterization of multi-species conserved sequences. Genome Res 13: 2507-2518.
162. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-1050.
163. Elnitski L, Hardison RC, Li J, Yang S, Kolbe D, et al. (2003) Distinguishing regulatory DNA from neutral sites. Genome Res 13: 64-72.
164. Kolbe D, Taylor J, Elnitski L, Eswara P, Li J, et al. (2004) Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. Genome Res 14: 700-707.
165. Siepel A, Pollard KS, Haussler D (2006) New methods for detecting lineage-specific selection. In Proc 10th Int'l Conf on Research in Computational Molecular Biology (RECOMB '06).
166. King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, et al. (2005) Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome Res 15: 1051-1060.
167. Sandelin A, Wasserman WW, Lenhard B (2004) ConSite: web-based prediction of regulatory elements using cross-species comparison. Nucleic Acids Res 32: W249-252.
168. Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. Genome Res 12: 832-839.
169. Corcoran DL, Feingold E, Dominick J, Wright M, Harnaha J, et al. (2005) Footer: a quantitative comparative genomics method for efficient recognition of cis-regulatory elements. Genome Res 15: 840-847.
170. Bigelow HR, Wenick AS, Wong A, Hobert O (2004) CisOrtho: a program pipeline for genome-wide identification of transcription factor target genes using phylogenetic footprinting. BMC Bioinformatics 5: 27.
171. Berezikov E, Guryev V, Cuppen E (2005) CONREAL web server: identification and visualization of conserved transcription factor binding sites. Nucleic Acids Res 33: W447-450.
172. Blanchette M, Tompa M (2003) FootPrinter: A program designed for phylogenetic footprinting. Nucleic Acids Res 31: 3840-3842.
173. Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. Genome Biol 5: R98.

174. Wakefield MJ, Maxwell P, Huttley GA (2005) Vestige: maximum likelihood phylogenetic footprinting. BMC Bioinformatics 6: 130.
175. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680.
176. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, et al. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 13: 721-731.
177. Pollard DA, Bergman CM, Stoye J, Celniker SE, Eisen MB (2004) Benchmarking tools for the alignment of functional noncoding DNA. BMC Bioinformatics 5: 6.
178. Rosenberg MS (2005) Multiple sequence alignment accuracy and evolutionary distance estimation. BMC Bioinformatics 6: 278.
179. Rosenberg MS (2005) Evolutionary distance estimation and fidelity of pair wise sequence alignment. BMC Bioinformatics 6: 102.
180. Michael M, Dieterich C, Vingron M (2005) SITEBLAST--rapid and sensitive local alignment of genomic sequences employing motif anchors. Bioinformatics 21: 2093-2094.
181. Blanco E, Messeguer X, Smith TF, Guigo R (2006) Transcription factor map alignment of promoter regions. PLoS Comput Biol 2: e49.
182. Dermitzakis ET, Clark AG (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol 19: 1114-1121.
183. Levy S, Hannenhalli S (2002) Identification of transcription factor binding sites in the human genome sequence. Mamm Genome 13: 510-514.
184. Ludwig MZ, Bergman C, Patel NH, Kreitman M (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403: 564-567.
185. Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, et al. (2005) Functional evolution of a cis-regulatory module. PLoS Biol 3: e93.
186. Scemama JL, Hunter M, McCallum J, Prince V, Stellwag E (2002) Evolutionary divergence of vertebrate Hoxb2 expression patterns and transcriptional regulatory loci. J Exp Zool 294: 285-299.
187. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, et al. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 301: 71-76.
188. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, et al. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299: 1391-1394.
189. Ovcharenko I, Boffelli D, Loots GG (2004) eShadow: a tool for comparing closely related sequences. Genome Res 14: 1191-1198.
190. Keightley PD, Lercher MJ, Eyre-Walker A (2005) Evidence for widespread degradation of gene control regions in hominid genomes. PLoS Biol 3: e42.
191. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, et al. (1991) Light-directed, spatially addressable parallel chemical synthesis. Science 251: 767-773.
192. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270: 484-487.

193. Bryan J (2004) Problems in gene clustering based on gene expression data. Journal of Multivariate Analysis 90: 44-66.
194. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11: 4241-4257.
195. Rhodes DR, Chinnaiyan AM (2005) Integrative analysis of the cancer transcriptome. Nat Genet 37 Suppl: S31-37.
196. Werner T (2001) Cluster analysis and promoter modelling as bioinformatics tools for the identification of target genes from expression array data. Pharmacogenomics 2: 25-36.
197. Ho Sui SJ, Mortimer JR, Arenillas DJ, Brumm J, Walsh CJ, et al. (2005) oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. Nucleic Acids Res 33: 3154-3164.
198. Haverty PM, Frith MC, Weng Z (2004) CARRIE web service: automated transcriptional regulatory network inference and interactive analysis. Nucleic Acids Res 32: W213-216.
199. Haverty PM, Hansen U, Weng Z (2004) Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. Nucleic Acids Res 32: 179-188.
200. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.
201. Long F, Liu H, Hahn C, Sumazin P, Zhang MQ, et al. (2004) Genome-wide prediction and analysis of function-specific transcription factor binding sites. In Silico Biol 4: 395-410.
202. Cora D, Di Cunto F, Provero P, Silengo L, Caselle M (2004) Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrepresented upstream motifs. BMC Bioinformatics 5: 57.
203. McNutt MC, Tongbai R, Cui W, Collins I, Freebern WJ, et al. (2005) Human promoter genomic composition demonstrates non-random groupings that reflect general cellular function. BMC Bioinformatics 6: 259.
204. Kelso J, Visagie J, Theiler G, Christoffels A, Bardien S, et al. (2003) eVOC: a controlled vocabulary for unifying gene expression data. Genome Res 13: 1222-1230.
205. OBO (2006) http://obo.sourceforge.net/.
206. Hurst LD, Pal C, Lercher MJ (2004) The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 5: 299-310.
207. Semon M, Duret L (2006) Evolutionary Origin and Maintenance of Coexpressed Gene Clusters in Mammals. Mol Biol Evol.
208. Pauli F, Liu Y, Kim YA, Chen PJ, Kim SK (2006) Chromosomal clustering and GATA transcriptional regulation of intestine-expressed genes in C. elegans. Development 133: 287-295.
209. Roy PJ, Stuart JM, Lund J, Kim SK (2002) Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans. Nature 418: 975-979.

210. Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, et al. (2004) An abundance of bidirectional promoters in the human genome. Genome Res 14: 62-66.
211. Rauscher FJ, 3rd, Voulalas PJ, Franza BR, Jr., Curran T (1988) Fos and Jun bind cooperatively to the AP-1 site: reconstitution in vitro. Genes Dev 2: 1687-1699.
212. Lin YS, Carey M, Ptashne M, Green MR (1990) How different eukaryotic transcriptional activators can cooperate promiscuously. Nature 345: 359-361.
213. Johnson AD (1995) The price of repression. Cell 81: 655-658.
214. Carey M, Lin YS, Green MR, Ptashne M (1990) A mechanism for synergistic activation of a mammalian gene by GAL4 derivatives. Nature 345: 361-364.
215. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431: 99-104.
216. Prestridge DS (1995) Predicting Pol II promoter sequences using transcription factor binding sites. J Mol Biol 249: 923-932.
217. Kondrakhin YV, Kel AE, Kolchanov NA, Romashchenko AG, Milanesi L (1995) Eukaryotic promoter recognition by binding sites for transcription factors. Comput Appl Biosci 11: 477-488.
218. Crowley EM, Roeder K, Bina M (1997) A statistical model for locating regulatory regions in genomic DNA. J Mol Biol 268: 8-14.
219. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, et al. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci U S A 99: 757-762.
220. Wasserman WW, Fickett JW (1998) Identification of regulatory regions which confer muscle-specific gene expression. J Mol Biol 278: 167-181.
221. Krivan W, Wasserman WW (2001) A predictive model for regulatory sequences directing liver-specific transcription. Genome Res 11: 1559-1566.
222. Wagner A (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. Bioinformatics 15: 776-784.
223. Frith MC, Hansen U, Weng Z (2001) Detection of cis-element clusters in higher eukaryotic DNA. Bioinformatics 17: 878-889.
224. Zhu Z, Shendure J, Church GM (2005) Discovering functional transcription-factor combinations in the human cell cycle. Genome Res 15: 848-855.
225. Ochoa-Espinosa A, Yucel G, Kaplan L, Pare A, Pura N, et al. (2005) The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. Proc Natl Acad Sci U S A 102: 4960-4965.
226. Markstein M, Markstein P, Markstein V, Levine MS (2002) Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. Proc Natl Acad Sci U S A 99: 763-768.
227. Kim JD, Hinz AK, Bergmann A, Huang JM, Ovcharenko I, et al. (2006) Identification of clustered YY1 binding sites in imprinting control region. Genome Res.
228. Frith MC, Spouge JL, Hansen U, Weng Z (2002) Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. Nucleic Acids Res 30: 3214-3224.

229. Frith MC, Li MC, Weng Z (2003) Cluster-Buster: Finding dense clusters of motifs in DNA sequences. Nucleic Acids Res 31: 3666-3668.
230. Alkema WB, Johansson O, Lagergren J, Wasserman WW (2004) MSCAN: identification of functional clusters of transcription factor binding sites. Nucleic Acids Res 32: W195-198.
231. Yu H, Yoo AS, Greenwald I (2004) Cluster Analyzer for Transcription Sites (CATS): a C++-based program for identifying clustered transcription factor binding sites. Bioinformatics 20: 1198-1200.
232. Donaldson IJ, Chapman M, Gottgens B (2005) TFBScluster: a resource for the characterization of transcriptional regulatory networks. Bioinformatics 21: 3058-3059.
233. Kel-Margoulis OV, Kel AE, Reuter I, Deineko IV, Wingender E (2002) TRANSCompel: a database on composite regulatory elements in eukaryotic genes. Nucleic Acids Res 30: 332-334.
234. Hochschild A, Ptashne M (1986) Cooperative binding of lambda repressors to sites separated by integral turns of the DNA helix. Cell 44: 681-687.
235. Fickett JW (1996) Coordinate positioning of MEF2 and myogenin binding sites. Gene 172: GC19-32.
236. Makeev VJ, Lifanov AP, Nazina AG, Papatsenko DA (2003) Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. Nucleic Acids Res 31: 6016-6026.
237. Kim B, Little JW (1992) Dimerization of a specific DNA-binding protein on the DNA. Science 255: 203-206.
238. Favorov AV, Gelfand MS, Gerasimova AV, Ravcheev DA, Mironov AA, et al. (2005) A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. Bioinformatics 21: 2240-2245.
239. Wagner A (1998) Distribution of transcription factor binding sites in the yeast genome suggests abundance of coordinately regulated genes. Genomics 50: 293-295.
240. Kozobay-Avraham L, Hosid S, Bolshoy A (2006) Involvement of DNA curvature in intergenic regions of prokaryotes. Nucleic Acids Res 34: 2316-2327.
241. Olivares-Zavaleta N, Jauregui R, Merino E (2006) Genome analysis of Escherichia coli promoter sequences evidences that DNA static curvature plays a more important role in gene transcription than has previously been anticipated. Genomics 87: 329-337.
242. Jerkovic B, Bolton PH (2000) The curvature of dA tracts is temperature dependent. Biochemistry 39: 12121-12127.
243. Bolshoy A, Nevo E (2000) Ecologic genomics of DNA: upstream bending in prokaryotic promoters. Genome Res 10: 1185-1193.
244. Ohyama T (2001) Intrinsic DNA bends: an organizer of local chromatin structure for transcription. Bioessays 23: 708-715.
245. Kiyama R, Trifonov EN (2002) What positions nucleosomes?--A model. FEBS Lett 523: 7-11.
246. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, et al. (2006) A genomic code for nucleosome positioning. Nature.

247. Espinosa-Urgel M, Tormo A (1993) Sigma s-dependent promoters in Escherichia coli are located in DNA regions with intrinsic curvature. Nucleic Acids Res 21: 3667-3670.
248. Carmona M, Magasanik B (1996) Activation of transcription at sigma 54-dependent promoters on linear templates requires intrinsic or induced bending of the DNA. J Mol Biol 261: 348-356.
249. Marilley M, Pasero P (1996) Common DNA structural features exhibited by eukaryotic ribosomal gene promoters. Nucleic Acids Res 24: 2204-2211.
250. Fitzgerald DJ, Dryden GL, Bronson EC, Williams JS, Anderson JN (1994) Conserved patterns of bending in satellite and nucleosome positioning DNA. J Biol Chem 269: 21303-21314.
251. Schatz T, Langowski J (1997) Curvature and sequence analysis of eukaryotic promoters. J Biomol Struct Dyn 15: 265-275.
252. Polak P, Domany E (2006) Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. BMC Genomics 7: 133.
253. Bannert N, Kurth R (2004) Retroelements and the human genome: new perspectives on an old relation. Proc Natl Acad Sci U S A 101 Suppl 2: 14572-14579.
254. Dunn CA, Romanish MT, Gutierrez LE, van de Lagemaat LN, Mager DL (2006) Transcription of two human genes from a bidirectional endogenous retrovirus promoter. Gene 366: 335-342.
255. Kashi Y, King DG (2006) Simple sequence repeats as advantageous mutators in evolution. Trends Genet 22: 253-259.
256. Fondon JW, 3rd, Garner HR (2004) Molecular origins of rapid and continuous morphological evolution. Proc Natl Acad Sci U S A 101: 18058-18063.
257. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, et al. (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15: 901-913.
258. Chiaromonte F, Yang S, Elnitski L, Yap VB, Miller W, et al. (2001) Association between divergence and interspersed repeats in mammalian noncoding genomic DNA. Proc Natl Acad Sci U S A 98: 14503-14508.
259. Aerts S, Van Loo P, Thijs G, Mayer H, de Martin R, et al. (2005) TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. Nucleic Acids Res 33: W393-396.
260. Hu Z, Fu Y, Halees AS, Kielbasa SM, Weng Z (2004) SeqVISTA: a new module of integrated computational tools for studying transcriptional regulation. Nucleic Acids Res 32: W235-241.
261. Edwards YJ, Carver TJ, Vavouri T, Frith M, Bishop MJ, et al. (2003) Theatre: A software tool for detailed comparative analysis and visualization of genomic sequence. Nucleic Acids Res 31: 3510-3517.
262. Gordon DB, Nekludova L, McCallum S, Fraenkel E (2005) TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. Bioinformatics 21: 3164-3165.
263. Che D, Jensen S, Cai L, Liu JS (2005) BEST: binding-site estimation suite of tools. Bioinformatics 21: 2909-2911.

264. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434: 338-345.
265. Chang LW, Nagarajan R, Magee JA, Milbrandt J, Stormo GD (2006) A systematic model to predict transcriptional regulatory mechanisms based on overrepresentation of transcription factor binding profiles. Genome Res 16: 405-413.
266. Wang T, Stormo GD (2005) Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. Proc Natl Acad Sci U S A 102: 17400-17405.
267. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.
268. Robertson G, Bilenky M, Lin K, He A, Yuen W, et al. (2006) cisRED: a database system for genome-scale computational discovery of regulatory elements. Nucleic Acids Res 34: D68-73.
269. Griffith OL, Pleasance ED, Fulton DL, Oveisi M, Ester M, et al. (2005) Assessment and integration of publicly available SAGE, cDNA microarray, and oligonucleotide microarray expression data for global coexpression analyses. Genomics 86: 476-488.
270. Blanchette M, Bataille AR, Chen X, Poitras C, Laganiere J, et al. (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res 16: 656-668.
271. Qin ZS, McCue LA, Thompson W, Mayerhofer L, Lawrence CE, et al. (2003) Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. Nat Biotechnol 21: 435-439.
272. Sharan R, Myers EW (2005) A motif-based framework for recognizing sequence families. Bioinformatics 21 Suppl 1: i387-393.
273. Gasch AP, Moses AM, Chiang DY, Fraser HB, Berardini M, et al. (2004) Conservation and evolution of cis-regulatory systems in ascomycete fungi. PLoS Biol 2: e398.
274. Snel B, van Noort V, Huynen MA (2004) Gene co-regulation is highly conserved in the evolution of eukaryotes and prokaryotes. Nucleic Acids Res 32: 4725-4731.
275. Wang T, Stormo GD (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. Bioinformatics 19: 2369-2380.
276. Johnson DS, Zhou Q, Yagi K, Satoh N, Wong W, et al. (2005) De novo discovery of a tissue-specific gene regulatory module in a chordate. Genome Res 15: 1315-1324.
277. Li N, Tompa M (2006) Analysis of computational approaches for motif discovery. Algorithms Mol Biol 1: 8.
278. Prakash A, Tompa M (2005) Discovery of regulatory elements in vertebrates through comparative genomics. Nat Biotechnol 23: 1249-1256.
279. Scherf M, Klingenhoff A, Werner T (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. J Mol Biol 297: 599-606.
280. Hahn MW, Stajich JE, Wray GA (2003) The effects of selection against spurious transcription factor binding sites. Mol Biol Evol 20: 901-906.
281. Weiss KM (1998) In search of human variation. Genome Res 8: 691-697.

282. Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Melmer G, et al. (1989) Identification of the cystic fibrosis gene: chromosome walking and jumping. Science 245: 1059-1065.
283. The Huntington's Disease Collaborative Research Group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. Cell 72: 971-983.
284. Crawford DC, Akey DT, Nickerson DA (2005) The patterns of natural variation in human genes. Annu Rev Genomics Hum Genet 6: 287-312.
285. Krawczak M, Reiss J, Cooper DN (1992) The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. Hum Genet 90: 41-54.
286. Blencowe BJ (2000) Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem Sci 25: 106-110.
287. Buckland PR, Coleman SL, Hoogendoorn B, Guy C, Smith SK, et al. (2004) A high proportion of chromosome 21 promoter polymorphisms influence transcriptional activity. Gene Expr 11: 233-239.
288. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928-933.
289. dbSNP (2006) http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi.
290. Ayala FJ, Escalante A, O'Huigin C, Klein J (1994) Molecular genetics of speciation and human origins. Proc Natl Acad Sci U S A 91: 6787-6794.
291. Zhivotovsky LA, Rosenberg NA, Feldman MW (2003) Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. Am J Hum Genet 72: 1171-1186.
292. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273: 1516-1517.
293. Collins FS, Guyer MS, Charkravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. Science 278: 1580-1581.
294. Lander ES (1996) The new genomics: global views of biology. Science 274: 536-539.
295. The HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299-1320.
296. Terwilliger JD, Hiekkalinna T (2006) An utter refutation of the "Fundamental Theorem of the HapMap". Eur J Hum Genet 14: 426-437.
297. Eichler EE (2006) Widening the spectrum of human genetic variation. Nat Genet 38: 9-11.
298. Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. Mol Ecol 11: 1-16.
299. Denoeud F, Vergnaud G, Benson G (2003) Predicting human minisatellite polymorphism. Genome Res 13: 856-867.
300. Marth GT (2003) Computational SNP discovery in DNA sequence data. Methods Mol Biol 212: 85-110.
301. Buetow KH, Edmonson MN, Cassidy AB (1999) Reliable identification of large numbers of candidate SNPs from public EST data. Nat Genet 21: 323-325.

302. Irizarry K, Kustanovich V, Li C, Brown N, Nelson S, et al. (2000) Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. Nat Genet 26: 233-236.
303. Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, et al. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407: 513-516.
304. Taillon-Miller P, Gu Z, Li Q, Hillier L, Kwok PY (1998) Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. Genome Res 8: 748-754.
305. Suh Y, Cantor C (2005) Single nucleotide polymorphisms (SNPs): detection, interpretation, and application. Mutat Res 573: 1-2.
306. Orita M, Iwahana H, Kanazawa H, Hayashi K, Sekiya T (1989) Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. Proc Natl Acad Sci U S A 86: 2766-2770.
307. Underhill PA, Jin L, Zemans R, Oefner PJ, Cavalli-Sforza LL (1996) A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. Proc Natl Acad Sci U S A 93: 196-200.
308. Sheffield VC, Cox DR, Lerman LS, Myers RM (1989) Attachment of a 40-base-pair G + C-rich sequence (GC-clamp) to genomic DNA fragments by the polymerase chain reaction results in improved detection of single-base changes. Proc Natl Acad Sci U S A 86: 232-236.
309. Myers RM, Larin Z, Maniatis T (1985) Detection of single base substitutions by ribonuclease cleavage at mismatches in RNA:DNA duplexes. Science 230: 1242-1246.
310. Chee M, Yang R, Hubbell E, Berno A, Huang XC, et al. (1996) Accessing genetic information with high-density DNA arrays. Science 274: 610-614.
311. Stanssens P, Zabeau M, Meersseman G, Remes G, Gansemans Y, et al. (2004) High-throughput MALDI-TOF discovery of genomic sequence polymorphisms. Genome Res 14: 126-133.
312. Dearlove AM (2002) High throughput genotyping technologies. Brief Funct Genomic Proteomic 1: 139-150.
313. Holland PM, Abramson RD, Watson R, Gelfand DH (1991) Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of Thermus aquaticus DNA polymerase. Proc Natl Acad Sci U S A 88: 7276-7280.
314. Ross P, Hall L, Smirnov I, Haff L (1998) High level multiplex genotyping by MALDI-TOF mass spectrometry. Nat Biotechnol 16: 1347-1351.
315. Ahmadian A, Gharizadeh B, Gustafsson AC, Sterky F, Nyren P, et al. (2000) Single-nucleotide polymorphism analysis by pyrosequencing. Anal Biochem 280: 103-110.
316. Wang DG, Fan JB, Siao CJ, Berno A, Young P, et al. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280: 1077-1082.
317. Shen R, Fan JB, Campbell D, Chang W, Chen J, et al. (2005) High-throughput SNP genotyping on universal bead arrays. Mutat Res 573: 70-82.
318. Sheet AGHMKD (2006) http://www.affymetrix.com/support/technical/datasheets/500k_datasheet.pdf.

319. Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, et al. (1999) A general approach to single-nucleotide polymorphism discovery. Nat Genet 23: 452-456.
320. Nickerson DA, Tobe VO, Taylor SL (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res 25: 2745-2751.
321. Stephens M, Sloan JS, Robertson PD, Scheet P, Nickerson DA (2006) Automating sequence-based detection and genotyping of SNPs from diploid samples. Nat Genet 38: 375-381.
322. Weckx S, Del-Favero J, Rademakers R, Claes L, Cruts M, et al. (2005) novoSNP, a novel computational tool for sequence variation discovery. Genome Res 15: 436-442.
323. Zhang J, Wheeler DA, Yakub I, Wei S, Sood R, et al. (2005) SNPdetector: A Software Tool for Sensitive and Accurate SNP Detection. PLoS Comput Biol 1: e53.
324. Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. Genome Res 11: 1725-1729.
325. Savage D, Batley J, Erwin T, Logan E, Love CG, et al. (2005) SNPServer: a real-time SNP discovery tool. Nucleic Acids Res 33: W493-495.
326. Unneberg P, Stromberg M, Sterky F (2005) SNP discovery using advanced algorithms and neural networks. Bioinformatics 21: 2528-2530.
327. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308-311.
328. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, et al. (2003) Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21: 577-581.
329. Fredman D, Munns G, Rios D, Sjoholm F, Siegfried M, et al. (2004) HGVbase: a curated resource describing human DNA variation and phenotype relationships. Nucleic Acids Res 32: D516-519.
330. Rajeevan H, Osier MV, Cheung KH, Deng H, Druskin L, et al. (2003) ALFRED: the ALelle FREquency Database. Update. Nucleic Acids Res 31: 270-271.
331. OMIM (2006) Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), June 2006. World Wide Web URL: http://www.ncbi.nlm.nih.gov/omim/.
332. Reumers J, Schymkowitz J, Ferkinghoff-Borg J, Stricher F, Serrano L, et al. (2005) SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. Nucleic Acids Res 33: D527-532.
333. Riva A, Kohane IS (2002) SNPper: retrieval and analysis of human SNPs. Bioinformatics 18: 1681-1685.
334. Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, et al. (2002) JSNP: a database of common gene variations in the Japanese population. Nucleic Acids Res 30: 158-162.
335. Tahira T, Baba S, Higasa K, Kukita Y, Suzuki Y, et al. (2005) dbQSNP: a database of SNPs in human promoter regions with allele frequency information determined by single-strand conformation polymorphism-based methods. Hum Mutat 26: 69-77.

336. Guryev V, Berezikov E, Cuppen E (2005) CASCAD: a database of annotated candidate single nucleotide polymorphisms associated with expressed sequences. BMC Genomics 6: 10.
337. Stitziel NO, Binkowski TA, Tseng YY, Kasif S, Liang J (2004) topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. Nucleic Acids Res 32: D520-522.
338. Ponomarenko JV, Merkulova TI, Orlova GV, Fokin ON, Gorshkova EV, et al. (2003) rSNP_Guide, a database system for analysis of transcription factor binding to DNA with variations: application to genome annotation. Nucleic Acids Res 31: 118-121.
339. SeattleSNPs (2006) NHLBI Program for Genomic Applications, SeattleSNPs, Seattle, WA (URL: http://pga.gs.washington.edu) [30 (Jul, 2006) accessed].
340. GeneSNPs (2006) NIEHS SNPs. NIEHS Environmental Genome Project, University of Washington, Seattle, WA (URL: http://egp.gs.washington.edu) [30 (Jul, 2006) accessed].
341. HSVD (2006) http://humanparalogy.gs.washington.edu/structuralvariation/.
342. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, et al. (2004) Detection of large-scale variation in the human genome. Nat Genet 36: 949-951.
343. Barber JC (2005) Directly transmitted unbalanced chromosome abnormalities and euchromatic variants. J Med Genet 42: 609-629.
344. Base BCMD (2006) http://research.nhgri.nih.gov/bic/.
345. Clifford R, Edmonson M, Hu Y, Nguyen C, Scherpbier T, et al. (2000) Expression-based genetic/physical maps of single-nucleotide polymorphisms identified by the cancer genome anatomy project. Genome Res 10: 1259-1265.
346. CFMDB (2006) http://www.genet.sickkids.on.ca/cftr/.
347. Mooney S (2005) Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. Brief Bioinform 6: 44-56.
348. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, et al. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22: 231-238.
349. Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics 158: 1227-1234.
350. Nguyen TH, Liu C, Gershon ES, McMahon FJ (2004) Frequency Finder: a multi-source web application for collection of public allele frequencies of SNP markers. Bioinformatics 20: 439-443.
351. Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155: 1405-1413.
352. Walsh EC, Sabeti P, Hutcheson HB, Fry B, Schaffner SF, et al. (2006) Searching for signals of evolutionary selection in 168 genes related to immune function. Hum Genet 119: 92-102.
353. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72.
354. Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, et al. (2006) Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet 38: 223-227.

355. Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11: 863-874.
356. Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30: 3894-3900.
357. Sunyaev S, Ramensky V, Bork P (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet 16: 198-200.
358. Krishnan VG, Westhead DR (2003) A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics 19: 2199-2209.
359. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31: 3812-3814.
360. Sunyaev S, Ramensky V, Koch I, Lathe W, 3rd, Kondrashov AS, et al. (2001) Prediction of deleterious human alleles. Hum Mol Genet 10: 591-597.
361. Yue P, Melamud E, Moult J (2006) SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics 7: 166.
362. Reumers J, Maurer-Stroh S, Schymkowitz J, Rousseau F (2006) SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non synonymous SNPs. Bioinformatics.
363. Chang H, Fujita T (2001) PicSNP: a browsable catalog of nonsynonymous single nucleotide polymorphisms in the human genome. Biochem Biophys Res Commun 287: 288-291.
364. Li JL, Li MX, Guo YF, Deng HY, Deng HW (2006) JADE: a distributed Java application for deleterious genomic mutation (DGM) estimation. Bioinformatics.
365. Mooney SD, Altman RB (2003) MutDB: annotating human variation with functionally relevant data. Bioinformatics 19: 1858-1860.
366. Freimuth RR, Stormo GD, McLeod HL (2005) PolyMAPr: programs for polymorphism database mining, annotation, and functional analysis. Hum Mutat 25: 110-117.
367. Conde L, Vaquerizas JM, Santoyo J, Al-Shahrour F, Ruiz-Llorente S, et al. (2004) PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. Nucleic Acids Res 32: W242-248.
368. Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3: 285-298.
369. Pastinen T, Hudson TJ (2004) Cis-acting regulatory variation in the human genome. Science 306: 647-650.
370. Knight JC (2005) Regulatory polymorphisms underlying complex disease traits. J Mol Med 83: 97-109.
371. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, et al. (2003) The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 20: 1377-1419.
372. Wittkopp PJ (2005) Genomic sources of regulatory variation in cis and in trans. Cell Mol Life Sci 62: 1779-1783.
373. Whitehead A, Crawford DL (2006) Variation within and among species in gene expression: raw material for evolution. Mol Ecol 15: 1197-1211.
374. King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. Science 188: 107-116.

375. Kornblihtt AR (2005) Promoter usage and alternative splicing. Curr Opin Cell Biol 17: 262-268.
376. Davidson EH (2001) Genomic regulatory systems : development and evolution. San Diego: Academic Press. xii, 261 p. p.
377. Rockman MV, Wray GA (2002) Abundant raw material for cis-regulatory evolution in humans. Mol Biol Evol 19: 1991-2004.
378. Pastinen T, Sladek R, Gurd S, Sammak A, Ge B, et al. (2004) A survey of genetic and epigenetic variation affecting human gene expression. Physiol Genomics 16: 184-193.
379. Knight JC, Keating BJ, Rockett KA, Kwiatkowski DP (2003) In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. Nat Genet 33: 469-475.
380. Hewett D, Lynch J, Child A, Firth H, Sykes B (1994) Differential allelic expression of a fibrillin gene (FBN1) in patients with Marfan syndrome. Am J Hum Genet 55: 447-452.
381. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, et al. (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430: 743-747.
382. Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, et al. (2004) Genetic inheritance of gene expression in human cell lines. Am J Hum Genet 75: 1094-1105.
383. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, et al. (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437: 1365-1369.
384. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, et al. (2005) Genome-wide associations of gene expression variation in humans. PLoS Genet 1: e78.
385. Deutsch S, Lyle R, Dermitzakis ET, Attar H, Subrahmanyan L, et al. (2005) Gene expression variation and expression quantitative trait mapping of human chromosome 21 genes. Hum Mol Genet 14: 3741-3749.
386. Pastinen T, Ge B, Hudson TJ (2006) Influence of human genome polymorphism on gene expression. Hum Mol Genet 15 Spec No 1: R9-16.
387. Tomso DJ, Inga A, Menendez D, Pittman GS, Campbell MR, et al. (2005) Functionally distinct polymorphic sequences in the human genome that are targets for p53 transactivation. Proc Natl Acad Sci U S A 102: 6431-6436.
388. Mottagui-Tabar S, Faghihi MA, Mizuno Y, Engstrom PG, Lenhard B, et al. (2005) Identification of functional SNPs in the 5-prime flanking sequences of human genes. BMC Genomics 6: 18.
389. Zhao T, Chang LW, McLeod HL, Stormo GD (2004) PromoLign: a database for upstream region analysis and SNPs. Hum Mutat 23: 534-539.
390. Khan IA, Mort M, Buckland PR, O'Donovan M C, Cooper DN, et al. (2005) In silico discrimination of single nucleotide polymorphisms and pathological mutations in human gene promoter regions by means of local DNA sequence context and regularity. In Silico Biol 6: 0003.
391. Wong GK, Yang Z, Passey DA, Kibukawa M, Paddock M, et al. (2003) A population threshold for functional polymorphisms. Genome Res 13: 1873-1879.

392. Ge B, Gurd S, Gaudin T, Dore C, Lepage P, et al. (2005) Survey of allelic expression using EST mining. Genome Res 15: 1584-1591.
393. Khaitovich P, Paabo S, Weiss G (2005) Toward a neutral evolutionary model of gene expression. Genetics 170: 929-939.
394. Balhoff JP, Wray GA (2005) Evolutionary analysis of the well characterized endo16 promoter reveals substantial variation within functional sites. Proc Natl Acad Sci U S A 102: 8591-8596.
395. Romano LA, Wray GA (2003) Conservation of Endo16 expression in sea urchins despite evolutionary divergence in both cis and trans-acting components of transcriptional regulation. Development 130: 4187-4199.
396. Tao H, Cox DR, Frazer KA (2006) Allele-specific KRT1 expression is a complex trait. PLoS Genet 2: e93.
397. Miao X, Yu C, Tan W, Xiong P, Liang G, et al. (2003) A functional polymorphism in the matrix metalloproteinase-2 gene promoter (-1306C/T) is associated with risk of development but not metastasis of gastric cardia adenocarcinoma. Cancer Res 63: 3987-3990.
398. Bond GL, Hu W, Bond EE, Robins H, Lutzker SG, et al. (2004) A single nucleotide polymorphism in the MDM2 promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans. Cell 119: 591-602.
399. Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, et al. (2003) Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. Science 301: 386-389.
400. Prokunina L, Castillejo-Lopez C, Oberg F, Gunnarsson I, Berg L, et al. (2002) A regulatory polymorphism in PDCD1 is associated with susceptibility to systemic lupus erythematosus in humans. Nat Genet 32: 666-669.
401. Kostrikis LG, Neumann AU, Thomson B, Korber BT, McHardy P, et al. (1999) A polymorphism in the regulatory region of the CC-chemokine receptor 5 gene influences perinatal transmission of human immunodeficiency virus type 1 to African-American infants. J Virol 73: 10264-10271.
402. Saito H, Tada S, Ebinuma H, Wakabayashi K, Takagi T, et al. (2001) Interferon regulatory factor 1 promoter polymorphism and response to type 1 interferon. J Cell Biochem 81: 191-200.
403. Montgomery SB, Astakhova T, Bilenky M, Birney E, Fu T, et al. (2004) Sockeye: a 3D environment for comparative genomics. Genome Res 14: 956-962.
404. Montgomery SB, Fu T, Guan J, Lin K, Jones SJ (2005) An application of peer-to-peer technology to the discovery, use and assessment of bioinformatics programs. Nat Methods 2: 563.
405. Montgomery SB, Griffith OL, Sleumer MC, Bergman CM, Bilenky M, et al. (2006) ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. Bioinformatics 22: 637-640.
406. Marra MA, Jones SJ, Astell CR, Holt RA, Brooks-Wilson A, et al. (2003) The Genome sequence of the SARS-associated coronavirus. Science 300: 1399-1404.
407. Krzywinski M, Montgomery SB, Jones SJ (2006) LADI - A Scientific Laboratory Network Document Management and Retrieval Tool Based on Google Desktop Search (in preparation).

408. Ashburner M, Drysdale R (1994) FlyBase--the Drosophila genetic database. Development 120: 2077-2079.
409. Stein L, Sternberg P, Durbin R, Thierry-Mieg J, Spieth J (2001) WormBase: network access to the genome and biology of Caenorhabditis elegans. Nucleic Acids Res 29: 82-86.
410. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, et al. (2000) PipMaker--a web server for aligning two genomic DNA sequences. Genome Res 10: 577-586.
411. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, et al. (2000) VISTA : visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16: 1046-1047.
412. Gottgens B, Gilbert JG, Barton LM, Grafham D, Rogers J, et al. (2001) Long-range comparison of human and mouse SCL loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. Genome Res 11: 87-97.
413. Lewis SE, Searle SM, Harris N, Gibson M, Lyer V, et al. (2002) Apollo: a sequence annotation editor. Genome Biol 3: RESEARCH0082.
414. Martz E (2002) Protein Explorer: easy yet powerful macromolecular visualization. Trends Biochem Sci 27: 107-109.
415. Sayle RA, Milner-White EJ (1995) RASMOL: biomolecular graphics for all. Trends Biochem Sci 20: 374.
416. Tate JG, Moreland JL, Bourne PE (2001) Design and implementation of a collaborative molecular graphics environment. J Mol Graph Model 19: 280-287, 369-273.
417. Kim SK, Lund JP, Kiraly M, Duke K, Jiang M, et al. (2001) A gene expression map for Caenorhabditis elegans. Science 293: 2087-2092.
418. Sensen CW (2002) Using CAVE technology for functional genomics studies. Diabetes Technol Ther 4: 867-871.
419. Bohannon J (2002) Bioinformatics. The human genome in 3D, at your fingertips. Science 298: 737.
420. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, et al. (2001) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 29: 11-16.
421. Maglott DR, Katz KS, Sicotte H, Pruitt KD (2000) NCBI's LocusLink and RefSeq. Nucleic Acids Res 28: 126-128.
422. Mangalam H (2002) The Bio* toolkits--a brief overview. Brief Bioinform 3: 296-302.
423. SARS C (2006) http://www.bcgsc.ca/bioinfo/SARS/.
424. Geisler JG, Stubbs LJ, Wasserman WW, Mucenski ML (1998) Molecular cloning of a novel mouse gene with predominant muscle and neural expression. Mamm Genome 9: 274-282.
425. Edwards YJ, Cottage A (2003) Bioinformatics methods to predict protein structure and function. A practical approach. Mol Biotechnol 23: 139-166.
426. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y (2003) Automatic prediction of protein function. Cell Mol Life Sci 60: 2637-2650.
427. Baker EN, Arcus VL, Lott JS (2003) Protein structure prediction and analysis as a tool for functional genomics. Appl Bioinformatics 2: S3-10.

428. Xiang Z, Yang Y, Ma X, Ding W (2003) Microarray expression profiling: analysis and applications. Curr Opin Drug Discov Devel 6: 384-395.
429. Conway T, Schoolnik GK (2003) Microarray expression profiling: capturing a genome-wide portrait of the transcriptome. Mol Microbiol 47: 879-889.
430. Quackenbush J (2001) Computational analysis of microarray data. Nat Rev Genet 2: 418-427.
431. Clifford RJ, Edmonson MN, Nguyen C, Scherpbier T, Hu Y, et al. (2004) Bioinformatics tools for single nucleotide polymorphism discovery and analysis. Ann N Y Acad Sci 1020: 101-109.
432. Fox JA, Butland SL, McMillan S, Campbell G, Ouellette BF (2005) The Bioinformatics Links Directory: a compilation of molecular biology web servers. Nucleic Acids Res 33: W3-24.
433. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16: 276-277.
434. Subramaniam S (1998) The Biology Workbench--a seamless database and analysis environment for the biologist. Proteins 32: 1-2.
435. Kapushesky M, Kemmeren P, Culhane AC, Durinck S, Ihmels J, et al. (2004) Expression Profiler: next generation--an online platform for analysis of microarray data. Nucleic Acids Res 32: W465-470.
436. Saeed AI, Sharov V, White J, Li J, Liang W, et al. (2003) TM4: a free, open-source system for microarray data management and analysis. Biotechniques 34: 374-378.
437. Wilkinson MD, Links M (2002) BioMOBY: an open source biological web services proposal. Brief Bioinform 3: 331-341.
438. Clark T, Martin S, Liefeld T (2004) Globally distributed object identification for biological knowledgebases. Brief Bioinform 5: 59-70.
439. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12: 1611-1618.
440. Kakazu KK, Cheung LW, Lynne W (2004) The Cancer Biomedical Informatics Grid (caBIG): pioneering an expansive network of information and tools for collaborative cancer research. Hawaii Med J 63: 273-275.
441. Hughes G, Mills H, De Roure D, Frey JG, Moreau L, et al. (2004) The semantic smart laboratory: a system for supporting the chemical eScientist. Org Biomol Chem 2: 3284-3293.
442. Finak G, Godin N, Hallett M, Pepin F, Rajabi Z, et al. (2005) BIAS: Bioinformatics Integrated Application Software. Bioinformatics 21: 1745-1746.
443. Eckart JD, Sobral BW (2003) A life scientist's gateway to distributed data management and computing: the PathPort/ToolBus framework. Omics 7: 79-88.
444. Teo YM, Wang X, Ng YK (2005) GLAD: a system for developing and deploying large-scale bioinformatics grid. Bioinformatics 21: 794-802.
445. Zuyderduyn SD, Jones SJ (2003) A knowledge discovery object model API for Java. BMC Bioinformatics 4: 51.
446. Letondal C (2001) A Web interface generator for molecular biology programs in Unix. Bioinformatics 17: 73-82.
447. Carver T, Bleasby A (2003) The design of Jemboss: a graphical user interface to EMBOSS. Bioinformatics 19: 1837-1843.

448. Basu MK (2001) SeWeR: a customizable and integrated dynamic HTML interface to bioinformatics services. Bioinformatics 17: 577-578.
449. Badidi E, De Sousa C, Lang BF, Burger G (2003) AnaBench: a Web/CORBA-based workbench for biomolecular sequence analysis. BMC Bioinformatics 4: 63.
450. Kerlavage A, Bonazzi V, di Tommaso M, Lawrence C, Li P, et al. (2002) The Celera Discovery System. Nucleic Acids Res 30: 129-136.
451. Stevens RD, Robinson AJ, Goble CA (2003) myGrid: personalised bioinformatics on the information grid. Bioinformatics 19 Suppl 1: i302-304.
452. Hoon S, Ratnapu KK, Chia JM, Kumarasamy B, Juguang X, et al. (2003) Biopipe: a flexible framework for protocol-based bioinformatics analysis. Genome Res 13: 1904-1915.
453. Oinn T, Addis M, Ferris J, Marvin D, Greenwood M, et al. (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics.
454. Shah SP, He DY, Sawkins JN, Druce JC, Quon G, et al. (2004) Pegasys: software for executing and integrating analyses of biological sequences. BMC Bioinformatics 5: 40.
455. Jareborg N, Durbin R (2000) Alfresco--a workbench for comparative genomic sequence analysis. Genome Res 10: 1148-1157.
456. Badidi E, Lang BF, Burger G (2004) FLOSYS--a web-accessible workflow system for protocol-driven biomolecular sequence analysis. Cell Mol Biol (Noisy-le-grand) 50: 785-793.
457. Qiao LA, Zhu J, Liu Q, Zhu T, Song C, et al. (2004) BOD: a customizable bioinformatics on demand system accommodating multiple steps and parallel tasks. Nucleic Acids Res 32: 4175-4181.
458. Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, et al. (2003) Toucan: deciphering the cis-regulatory logic of coregulated genes. Nucleic Acids Res 31: 1753-1764.
459. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, et al. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31: 3784-3788.
460. Plewniak F, Bianchetti L, Brelivet Y, Carles A, Chalmel F, et al. (2003) PipeAlign: A new toolkit for protein family analysis. Nucleic Acids Res 31: 3829-3832.
461. Holm L, Park J (2000) DaliLite workbench for protein structure comparison. Bioinformatics 16: 566-567.
462. Shah SP, McVicker GP, Mackworth AK, Rogic S, Ouellette BF (2003) GeneComber: combining outputs of gene prediction programs for improved results. Bioinformatics 19: 1296-1297.
463. Makalowska I, Ryan JF, Baxevanis AD (2001) GeneMachine: gene prediction and sequence annotation. Bioinformatics 17: 843-844.
464. Higa RH, Togawa RC, Montagner AJ, Palandrani JC, Okimoto IK, et al. (2004) STING Millennium Suite: integrated software for extensive analyses of 3d structures of proteins and their complexes. BMC Bioinformatics 5: 107.
465. Shatsky M, Dror O, Schneidman-Duhovny D, Nussinov R, Wolfson HJ (2004) BioInfo3D: a suite of tools for structural bioinformatics. Nucleic Acids Res 32: W503-507.

466. Moreland JL, Gramada A, Buzko OV, Zhang Q, Bourne PE (2005) The Molecular Biology Toolkit (MBT): a modular platform for developing molecular visualization applications. BMC Bioinformatics 6: 21.
467. Wildpaner M, Schneider G, Schleiffer A, Eisenhaber F (2001) Taxonomy workbench. Bioinformatics 17: 1179-1182.
468. Drummond A, Strimmer K (2001) PAL: an object-oriented programming library for molecular evolution and phylogenetics. Bioinformatics 17: 662-663.
469. Price EW, Carbone I (2005) SNAP: workbench management tool for evolutionary population genetic analysis. Bioinformatics 21: 402-404.
470. Choi K, Ma Y, Choi JH, Kim S (2005) PLATCOM: a Platform for Computational Comparative Genomics. Bioinformatics.
471. Elnitski L, Riemer C, Petrykowska H, Florea L, Schwartz S, et al. (2002) PipTools: a computational toolkit to annotate and analyze pairwise comparisons of genomic sequences. Genomics 80: 681-690.
472. Cheung KH, de Knikker R, Guo Y, Zhong G, Hager J, et al. (2004) Biosphere : the interoperation of web services in microarray cluster analysis. Appl Bioinformatics 3: 253-256.
473. Sauro HM, Hucka M, Finney A, Wellock C, Bolouri H, et al. (2003) Next generation simulation tools: the Systems Biology Workbench and BioSPICE integration. Omics 7: 355-372.
474. Maki Y, Takahashi Y, Arikawa Y, Watanabe S, Aoshima K, et al. (2004) An integrated comprehensive workbench for inferring genetic networks: voyagene. J Bioinform Comput Biol 2: 533-550.
475. Toyoda T, Konagaya A (2003) KnowledgeEditor: a new tool for interactive modeling and analyzing biological pathways based on microarray data. Bioinformatics 19: 433-434.
476. Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using Weka. Bioinformatics 20: 2479-2481.
477. Vaidyanathan G (2004) InfoEvolve: moving from data to knowledge using information theory and genetic algorithms. Ann N Y Acad Sci 1020: 227-238.
478. Documentation; C (2006) http://www.bcgsc.ca/gc/bomge/chinook/docs.
479. Java (2006) http://java.sun.com.
480. JXTA (2006) http://www.jxta.org.
481. Berezikov E, Guryev V, Plasterk RH, Cuppen E (2004) CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. Genome Res 14: 170-178.
482. Morgenstern B (2004) DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. Nucleic Acids Res 32: W33-36.
483. Darling AC, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14: 1394-1403.
484. Brudno M, Malde S, Poliakov A, Do CB, Couronne O, et al. (2003) Glocal alignment: finding rearrangements during alignment. Bioinformatics 19 Suppl 1: i54-62.
485. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302: 205-217.

486. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365-386.
487. Bailey TL, Elkan C (1995) The value of prior knowledge in discovering motifs with MEME. Proc Int Conf Intell Syst Mol Biol 3: 21-29.
488. van Helden J (2003) Regulatory sequence analysis tools. Nucleic Acids Res 31: 3593-3596.
489. Sinha S, van Nimwegen E, Siggia ED (2003) A probabilistic method to detect regulatory modules. Bioinformatics 19 Suppl 1: i292-301.
490. Hertz GZ, Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15: 563-577.
491. Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268: 78-94.
492. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res 8: 967-974.
493. Frith MC, Fu Y, Yu L, Chen JF, Hansen U, et al. (2004) Detection of functional DNA motifs via statistical over-representation. Nucleic Acids Res 32: 1372-1381.
494. Stein L (2002) Creating a bioinformatics nation. Nature 417: 119-120.
495. States DJ (2002) Bioinformatics code must enforce citation. Nature 417: 588.
496. Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, et al. (2004) An overview of Ensembl. Genome Res 14: 925-928.
497. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res 32: D91-94.
498. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2005) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 33: D39-45.
499. Maglott D, Ostell J, Pruitt KD, Tatusova T (2005) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 33: D54-58.
500. Ponomarenko JV, Merkulova TI, Vasiliev GV, Levashova ZB, Orlova GV, et al. (2001) rSNP_Guide, a database system for analysis of transcription factor binding to target sequences: application to SNPs and site-directed mutations. Nucleic Acids Res 29: 312-316.
501. Booth D, Haas, H., McCabe, F., Newcomer, E., Champion, M., Ferris, C., Orchard, D (2004) Web Services architecture. W3C working group note, W3C.
502. Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L (2001) The distributed annotation system. BMC Bioinformatics 2: 7.
503. RegCreative (2006) http://www.dmbr.ugent.be/bioit/contents/regcreative/.
504. Wang WY, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6: 109-118.
505. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6: 95-108.
506. Belmont JW, Gibbs RA (2004) Genome-wide linkage disequilibrium and haplotype maps. Am J Pharmacogenomics 4: 253-262.

507. Miller RD, Kwok PY (2001) The birth and death of human single-nucleotide polymorphisms: new experimental evidence and implications for human history and medicine. Hum Mol Genet 10: 2195-2198.
508. Sadee W, Dai Z (2005) Pharmacogenetics/genomics and personalized medicine. Hum Mol Genet 14 Spec No. 2: R207-214.
509. Buckland PR, Hoogendoorn B, Coleman SL, Guy CA, Smith SK, et al. (2005) Strong bias in the location of functional promoter polymorphisms. Hum Mutat 26: 214-223.
510. Goodsell DS, Dickerson RE (1994) Bending and curvature calculations in B-DNA. Nucleic Acids Res 22: 5497-5503.
511. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P (2004) Coexpression analysis of human genes across many microarray data sets. Genome Res 14: 1085-1094.
512. Noble WS, Kuehn S, Thurman R, Yu M, Stamatoyannopoulos J (2005) Predicting the in vivo signature of human gene regulatory sequences. Bioinformatics 21 Suppl 1: i338-343.
513. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, et al. (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 31: 3497-3500.
514. Pavlidis P, Wapinski I, Noble WS (2004) Support vector machine classification on the web. Bioinformatics 20: 586-587.
515. Di Rienzo A, Hudson RR (2005) An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends Genet 21: 596-601.
516. Cannata N, Merelli E, Altman RB (2005) Time to organize the bioinformatics resourceome. PLoS Comput Biol 1: e76.
517. Wilkinson M, Schoof H, Ernst R, Haase D (2005) BioMOBY successfully integrates distributed heterogeneous bioinformatics Web Services. The PlaNet exemplar case. Plant Physiol 138: 5-17.