Bioinformatics Tools for the Analysis of Gene-Phenotype Relationships Coupled with a Next Generation ChIP-Sequencing Data Processing Pipeline

Erinija Pranckeviciene

Thesis submitted to the

Faculty of Graduate and Postdoctoral Studies

in partial fulfillment of the requirements

for the Doctorate in Philosophy degree in Cellular and Molecular Medicine

Department of Cellular and Molecular Medicine

Faculty of Medicine

University of Ottawa

Abstract

The rapidly advancing high-throughput and next generation sequencing technologies facilitate deeper insights into the molecular mechanisms underlying the expression of phenotypes in living organisms. Experimental data and scientific publications following this technological advancement have rapidly accumulated in public databases. Meaningful analysis of currently available data in genomic databases requires sophisticated computational tools and algorithms, and presents considerable challenges to molecular biologists without specialized training in bioinformatics. To study their phenotype of interest molecular biologists must prioritize large lists of poorly characterized genes generated in high-throughput experiments. To date, prioritization tools have primarily been designed to work with phenotypes of human diseases as defined by the genes known to be associated with those diseases. There is therefore a need for more prioritization tools for phenotypes which are not related with diseases generally or diseases with which no genes have yet been associated in particular. Chromatin immunoprecipitation followed by next generation sequencing (ChIP-Seq) is a method of choice to study the gene regulation processes responsible for the expression of cellular phenotypes. Among publicly available computational pipelines for the processing of ChIP-Seq data, there is a lack of tools for the downstream analysis of composite motifs and preferred binding distances of the DNA binding proteins. This thesis is aimed to address the gap existing in the tools available to process high-throughput ChIP-Seq data to provide rapid analysis and interpretation of large lists of poorly characterized genes. Additionally, programs for the analysis of preferred binding distances of transcription factors were integrated into the pipeline for expedited results. A gene prioritization algorithm linking genes to non-disease phenotypes described by meaningful keywords was developed. This algorithm can be used to process candidate genetic targets of a transcription factor produced by a computational pipeline for ChIP-Seq data analysis.

Contents

List of Figures

List of Tables

Abbreviations

API	**A**pplication **P**rogramming **I**nterface
bp	**b**ase **pair**
BWA	**B**urrows **W**heeler **A**lgorithm
ChIP-Seq	**C**hromatin **I**mmunoprecipitation followed by **Seq**uencing
CRM	**C**is **R**egulatory **M**odule
CV	**C**ross **V**alidation
ECFCs	**E**ndothelial **C**olony-**F**orming **C**ells
ENCODE	The **ENC**yclopedia of **D**NA **E**lements
DNA	**D**eoxyribo**N**ucleic**A**cid
EQ	**E**ntity **Q**uality
FBR	**F**uzzy **B**inary **R**elationship
GAD	**G**enetic **A**ssociation **D**atabase
GO	**G**ene **O**ntology
GSEA	**G**ene **S**et **E**nrichment **A**nalysis
GWAS	**G**enome **W**ide **A**ssociation **S**tudy
HPC	**H**igh **P**erformance **C**omputing
HT	**H**igh **T**hroughput
IC	**I**nformation **C**ontent
Kb	**K**ilo **b**ases
LOF	**L**oss **O**f **F**unction
NCBI	**N**ational **C**enter of **B**iotechnology **I**nformation
NGS	**N**ext **G**eneration **S**equencing
NLM	**N**ational **L**ibrary of **M**edicine
MAQ	**M**apping and **A**lignment with **Q**uality scores
MEA	**M**odular **E**nrichment **A**nalysis
MeSH	**Me**dical **S**ubject **H**eadings

MGI	**M**ouse **G**enome **I**nformatics
MPO	**M**ammalian **P**henotype **O**ntology
OBO	**O**pen **B**iomedical **O**ntology
OMIM	**O**nline **M**endelian **I**nformation for **M**an
OWL	**O**ntology **W**eb **L**anguage
PCR	**P**olymerase **C**hain **R**eaction
PFM	**P**osition **F**requency **M**atrix
PMID	**P**ub**M**ed **ID**entifier
PWM	**P**osition **W**eight **M**atrix
RDF	**R**esource **D**escription **F**ramework
RNA	**R**ibo**n**ucleic **A**cid
RT-qPCR	**R**real **T**ime **q**uantitative **P**olymerase **C**hain **R**eaction
SEA	**S**ingular **E**nrichment **A**nalysis
SGE	**S**un **G**rid **E**ngine
SNP	**S**ingle **N**ucleotide **P**osition
T ALL	**A**cute **T** **L**ymphoblastic **L**eukemia
TF	**T**ranscription **F**actor
TSS	**T**ranscription **S**tart **S**ite
UCSC	**U**niversity of **C**alifornia **S**anta **C**ruz
UTR	**U**n**T**ranslated **R**egion

Acknowledgements

This work could not have been completed without the help and support of many people.

I am grateful to my advisors Dr. Carol Perez-Iratxeta, Dr. Jeffrey Dilworth and Dr. Marjorie Brand for providing the support and opportunity to work on this exciting project and to be able to make a meaningful contribution to the constantly changing field of bioinformatics.

I thank my Thesis Advisory Committee members Dr. Alexandre Blais, Dr. Harold Atkins and Dr. Duncan Stewart for their guidance and advice.

The director of the Cellular and Molecular Medicine graduate program, Dr. Nadine Wiper-Bergeron, and the graduate program administrative assistant, Mrs. Sylvie Deblois, are greatly acknowledged for their kind help in all matters related to advancing me through the graduate program.

The University of Ottawa is gratefully acknowledged for their Admission Scholarship.

The Bioinformatics Team at the Ottawa Hospital Research Institute helped me overcome difficult technical challenges. I thank Mr. Gareth Palidwor, Mr. Christopher Porter, Dr. Parameswaran Ramachandran and Ms. Caroline Luis-Jeune for that. I would especially like to acknowledge the indispensable help and encouragement of Gareth and Caroline in building the web application for my project, for professional conversations and also for being colleagues and friends.

I greatly acknowledge the Ottawa Hospital Research Institute system administrator, Mr. Alan McKay, for all of his help in data management and server side programming.

I thank the current and past members of the Dr. Dilworth and Dr. Brand Laboratories for being so friendly to me and sharing their molecular biology expertise with me. For this I especially thank Dr. Carmen Palii, Dr. Prakash Chaturvedi, Ms. Brinda Somasundaran, Ms. Amanda Dai, Dr. Herve Faralli, Dr. Arif Aziz and Mrs. LiFang Li.

My Ottawa University friends Ana van Weringh, Manon Ragonnet, Oksana Korol, Hilary Phenix and Nezar Abdennur have always inspired me. I thank them all for their friendship and creative

discussions.

I thank Dr. Arezu Ashl-Jahani for her insightful professional communications.

I gratefully acknowledge the kind help and support of Ms. Thelma Claire Wilson in editing and proofreading my written work.

I thank all whom I interacted with and learned from during this stage of my professional life. I thank Mr. Mandeep Metharu for inspiring communications. I thank Ms. Lindsey Li for her help and suggestions with respect to text editing.

I am immensely grateful for the patience, understanding, love and support of my two sons, Matas and Jonas, and my parents, Ramute and Leo.

I thank Fr. Roman Rytsar for his spiritual guidance while I was writing my thesis.

Preface

This thesis addresses the needs of applied bioinformatics research. It is a result of the work I performed as a part of the two research projects I completed during my PhD training at the Ottawa Hospital Research Institute.

The first research project, conducted in the laboratory of Dr. J. Dilworth, concerned the regulation of myoblast cell fusion. In the context of this research project I was developing an algorithm to prioritize differentially-expressed genes with respect to the cell fusion phenotype. Early results of this research were used in the publication **Comparative expression profiling identifies differential roles for Myogenin and p38-MAPK signaling in myogenesis** [Liu et al., 2012], to which I contributed as a co-author.

In the laboratory of Dr. M. Brand I developed a computational pipeline for processing data obtained via Chromatin immunoprecipitation and massively parallel sequencing (ChIP-Seq). In addition to the pipeline, I developed a novel tool which identifies the distances at which two transcription factors (TF) will prefer to co-occur in the predicted TF binding regions. The developed pipeline and new TF tool were used in the co-authored publication **Trichostatin A enhances vascular repair by injected human endothelial progenitors through increasing the expression of TAL1-dependent genes** [Palii et al., 2014].

The contribution of this Thesis is my gene prioritization algorithm, which I developed in the first project, and the ChIP-Seq data analysis pipeline, which made use of the TF tool I created in the second research project. A major part of this Thesis consists of the conceptual development and validation of my algorithm in performing gene prioritization with respect to the general phenotypes defined by the keywords used.

Chapter 1

Introduction

The expression of a phenotype or disease starts at the molecular level, evolves into the cellular and tissue levels, and finally manifests in the organism. Understanding the molecular mechanisms underlying both normal cellular processes and disease-causing gene perturbations has numerous applications in clinical diagnostics, personal genomics and engineering [Dudley and Karczewski, 2013, Esvelt and Wang, 2013, Fernald et al., 2011, Lee et al., 2013, Shendure and Lieberman Aiden, 2012]. High-throughput (HT) technologies to acquire genomic data enable deep insights into the molecular mechanisms of living systems, and are rapidly advancing genomic studies. Most of these studies address two major questions in one way or another:

What genomic and molecular markers are associated with an observed phenotype?

What molecular mechanisms lead to that phenotype in the studied organism?

Answering these questions and uncovering gene-phenotype relationships relies on experimental research that has already generated very large amounts of high-throughput data stored in public databases [Aerts et al., 2009, Carpenter and Sabatini, 2004, de la Iglesia D et al., 2013, Dunham et al., 2012, Hawkins et al., 2010]. New knowledge about genes and their functions is acquired all the time based on the constant gathering of genomic data. To date there are about 1500 databases hosting various types of genomic and molecular biology data [Fernandez-Suarez and

Galperin, 2013]. The number of research publications analyzing newly-generated data is rapidly increasing as well [Manconi et al., 2012]. Vast information resources consisting of publicly available literature and high-throughput microarray and sequencing data of many organisms can be utilized in the search for genetic factors affecting phenotype expressions [Baxevanis, 2011, Tiffin et al., 2009]. However, the true value of these data in discoveries of gene-to-phenotype relationships depends on our ability to make correct links between molecular entities and the information stored in databases and literature [de la Iglesia D et al., 2013, Hawkins et al., 2010].

For this reason integrative algorithms to analyze high-throughput data by mining genomic databases and literature have been the focus of intensive research resulting in many publicly-available bioinformatics tools intended for biologists and clinical researchers [Andronis et al., 2011, Brazas et al., 2012, de la Iglesia D et al., 2013, Kersey and Apweiler, 2006, Krallinger et al., 2010, Rebholz-Schuhmann et al., 2008, Turenne et al., 2012, Zhu et al., 2010b]. The stages

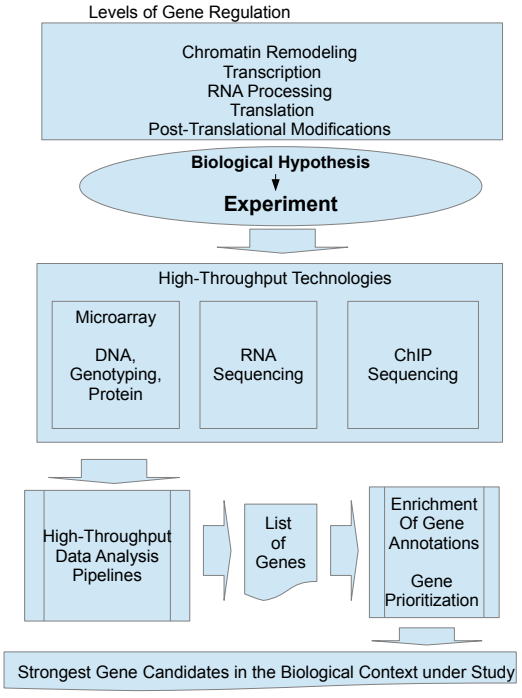


FIGURE 1.1: Stages of High-Throughput Genomic Data Analysis

of genomic data analysis are illustrated in Figure 1.1. At first high-throughput data is collected

experimentally to test a biological hypothesis. Different types of HT data capture different aspects of biological processes. Examples of HT data types are shown in Table 1.1. This raw data

TABLE 1.1: High-Throughput Technologies for the Profiling of Genomic Data

Description	**Scope of Application**
ChIP Sequencing Technology	
Sites of DNA interaction with protein are isolated by chromatin immunoprecipitation and produces a library of DNA sites bound to protein in vivo. This library is subjected to massively parallel sequencing.	Measures protein-DNA interactions genome-wide and identifies TFBS and chromatin modifications.
DNA Microarray Technology	
DNA chips immobilized with probes to match the florescent-labeled DNA in the biological specimen. The abundance of the resulting hybridization is measured by optically scanning the microarray chip. It has limited coverage. Several categories of microarrays exist: cDNA, oligonucleotide.	Comparison of gene expression under specific biological conditions; genotyping.
RNA Sequencing Technology	
Sequencing of whole cell transcriptome. Measures abundance of RNA transcripts including miRNA, tRNA, ribosomal profiling.	Facilitates measuring of gene expression, alternative splicing, exon-intron junctions,post-transcriptional changes, gene mutations.

is processed by computational pipelines usually resulting in large lists of genes. These lists are further analyzed through integrative computational methods to dissect individual or collective gene involvement in the biological function that is being investigated. Very often the genes in the list are poorly characterized. This presents considerable challenges in clarifying the particular roles of genes and their functions since it is almost impossible to experimentally test all the individual and combined effects of the genomic factors in a large list [Tranchevent, 2011]. Large numbers of experiments have considerable costs; therefore, computational algorithms are used to select a small number of the genomic factors for followup. Tasks, in which a few genes that are strongly linked to the phenotype of interest have to be selected from the identified list of genes, are encountered quite often:

- **A selection of a subset of differentially-expressed genes identified in a microarray or RNA-Seq experiment**. Example: A microarray experiment has been performed to identify genes whose expression is required for cell fusion to take place in a cell [Liu et al., 2012]. 214 diferentially-expressed genes were identified. Which of those genes are most related to the cell fusion phenotype?

- **Uncovering relationships between a biological process of interest and target genes regulated by a transcription factor in a ChIP-Seq experiment**. Example: A ChIP-Seq experiment has been performed to identify genome-wide binding regions of the TAL1 transcription factor in jurkat and erythroid cells in order to identify the target genes of TAL1 in those cells [Palii et al., 2011]. The gene lists consist of identified target genes in the two cell lines. Which of those genes on these lists are associated with leukemia?

- **Finding a causative gene harbored in a chromosomal deletion or duplication identified in a patient with a clinical phenotype**. Example: A microduplication has been identified in a patient with an intellectual disability phenotype [Preiksaitiene et al., 2012]. The list comprises genes present in the identified microduplication. Which of these genes are most associated with the clinical phenotype observed in the patient?

- **Finding candidate genes in gene lists collected from literature**. Example: Some gene lists reported in the literature are associated with definite biological processes through experimental evidence and are stored in databases such as the database of molecular signatures, MSigDB. Which genes in the selected lists of interest might be linked to a phenotype other than the associated one? What other functions are these genes associated with?

Two independent stages can be distinguished in searching for a gene-phenotype relationship : (i) obtaining a list of genes; (ii) elucidating a few candidate genes in the list that are strongly related to the phenotype of interest. Methods used in the first stage - to obtain the list - depend on

the modality of the data (microarray, ChIP-Seq, literature or else). Methods of the second stage - to elucidate a few best candidates - usually apply analyses of the enrichment of annotations and gene prioritization. In the second stage the chosen analysis strategy is determined by the particular needs of the researcher [Huang et al., 2007, Moreau and Tranchevent, 2012].

This thesis is concerned with a set of tools to perform analysis in both stages. A computational pipeline was developed to perform **ChIP-Seq** data analysis leading to a list of genes in the first stage. In the second stage, that of gene prioritization, a strategy was developed to infer relationships between genes and phenotypes, emphasizing non-disease phenotypes. In what follows, state-of-the-art approaches in ChIP-Seq data analysis and existing strategies to infer gene-phenotype relationships for the purposes of gene prioritization are reviewed, so as to demonstrate the merit of the new tools I developed for this thesis.

1.1 ChIP-Seq Data Analysis

DNA-binding proteins and their complexes have important regulatory roles in cellular processes such as transcription, chromatin remodeling and translation [Boyle et al., 2011, Pike, 2011]. Chromatin immunoprecipitation followed by next generation sequencing (ChIP-Seq) has become a standard technique to analyze DNA-binding proteins by identifying their binding locations or modified histones genome-wide [Furey, 2012, Pepke et al., 2009]. The ChIP-Seq experiment has the following steps. The protein of interest is crosslinked with the DNA it binds to in-vitro. The resulting chromatin fragments are broken up by sonication. Those bound by the protein of interest are then selected by immunoprecipitation: the protein of interest, which is still cross-linked to DNA fragments, is selected by means of antibodies against that protein. The DNA is then separated from the protein and subjected to amplification and massively parallel sequencing. These steps of Chip-Seq experiments are shown in Figure 1.2.

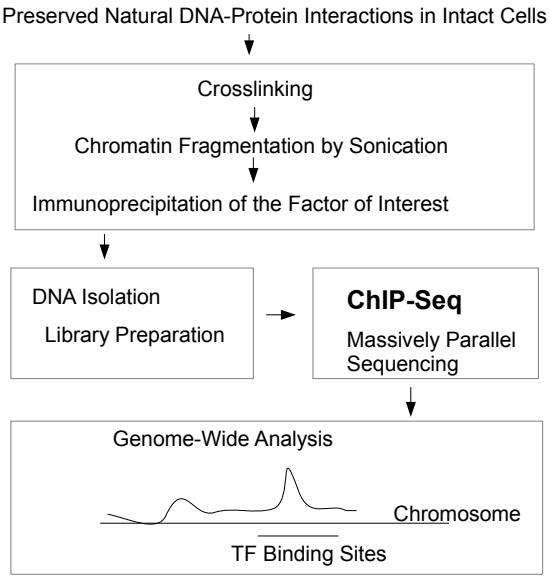


FIGURE 1.2: Steps of ChIP-Seq Experiment

Usually the raw ChIP-Seq data consists of millions of reads. The sizes of the reads vary by instrument (26-150+ base pairs) [Pepke et al., 2009]. These raw reads obtained in sequencing are computationally processed to identify a genome-wide list of protein binding sites and their regulated genes [Kharchenko et al., 2008, Landt et al., 2012]. Computational procedures performed on the short reads to obtain a list of the target genes of a transcription factor are surveyed in the following section 1.1.1. A brief summary of these computational steps and the software most frequently used in ChIP-Seq data analysis is shown in Table 1.2.

1.1.1 Computational Processing of Sequenced Reads

Quality Control. At first the reads have to pass through quality control filters [Patel and Jain, 2012]. Poor-quality reads [Cock et al., 2010] and reads contaminated with adapters during the library preparation step are removed from the data prior to mapping [Koboldt et al., 2010, Leleu et al., 2010]. In order to preserve as much information as possible, the contaminated reads can be trimmed and re-mapped. The reads should not be trimmed very short (less than 26 bp)

TABLE 1.2: ChIP-Seq Data Processing Steps

Tasks	**Frequently Used Software**
1. Quality Control	
Quality report. Removal of bad quality reads, removing or trimming the reads contaminated by the linkers and adapters, identifying duplicates.	FASTQC [Schmieder and Edwards, 2011], clean reads [Blanca et al., 2011], SeqTrim [Falgueras et al., 2010].
2. Alignment to the Reference Genome	
The short reads are mapped to the reference genome.	MAQ [Li et al., 2008], BWA [Li and Durbin, 2009], Bowtie [Langmead et al., 2009].
3. Identification of Transcription Factor Binding Sites Genome-Wide	
Evaluation of the enrichment of the short read tags over the genome and identification of the peaks. The enriched regions/peaks contain putative TF binding sites. Estimation of the sufficiency of the coverage of the identified genomic regions by the short reads.	MACS [Zhang et al., 2008], PeakRanger [Feng et al., 2011b], SISSRs [Jothi et al., 2008], BayesPeak [Cairns et al., 2011], Bioconductor packages [Morgan et al., 2009].
4. Motif Analysis	
Scanning the identified genomic regions for the presence of the known DNA-binding sequences. Identification of *de novo* motifs. Identification of the preferred binding distances of the motifs.	MEME [Bailey et al., 2009], DREME [Bailey, 2011], HOMER [Heinz et al., 2010], Bioconductor package MotifRG [Fong et al., 2012], rGADEM [Mercier et al., 2011].
5. Annotation of Genomic Regions	
Annotation of peaks by the genomic features: promoters, untranslated regions, exons, introns. Summary of the annotations. Investigation of the peak distances with respect to gene transcription start sites (TSS).	HOMER [Heinz et al., 2010], ChIPseeker [Giannopoulou and Elemento, 2011], ChIPpeakAnno [Zhu et al., 2010a].
6. Peak-Gene Association and Enrichment of Annotations of Associated Genes	
Identification of the closest genes in given ranges of genomic distances. Identification of putative regulatory relationships between the TF and the identified genes. Analysis of the functional enrichment of the genes.	BedTools [Quinlan and Hall, 2010], GREAT [McLean et al., 2010], ChIPseeker [Giannopoulou and Elemento, 2011].

since such reads may have multiple matches in the reference genome, introducing unnecessary ambiguity into the resulting mapping [Schmieder and Edwards, 2011].

Mapping to the reference genome. The short reads passing the quality control step are aligned to the reference genome of the organism under study. The alignment requires significant

computing power and sophisticated algorithms because it can be complicated by sequencing errors [Li and Homer, 2010]. The first widely used program that addressed those challenges was MAQ [Li et al., 2008]. It considered the base quality scores during sequence alignment which helped address the variable quality of a sequence across a read. The programs that followed MAQ were BOWTIE [Langmead et al., 2009], BWA [Li and Durbin, 2009] and SOAP2 [Li et al., 2009]. By using the Burrows-Wheeler transformation algorithm they dramatically decreased the mapping time [Li and Durbin, 2009, 2010] and became standard aligners. Faster mapping algorithms for gapped alignments are still being actively developed [Hatem et al., 2013, Ruffalo et al., 2011].

Visualization. To investigate the locations at which the ChIP-Seq data have a strong signal, the aligned data is displayed in an integrative genome visualization environment. The UCSC Genome Browser is one of the best environments to date because it facilitates analysis of the ChIP-Seq signals in various biological contexts [Karolchik et al., 2012]. It integrates all available data about the genomic elements from the Encyclopedia of DNA Elements (ENCODE) project [Rosenbloom et al., 2012], [Dunham et al., 2012, Hoffman et al., 2013]. It hosts tracks of chromatin methylation profiles, gene conservation information, and phenotype information from the Online Mendelian Information for Man (OMIM) database [Karolchik et al., 2012, Meyer et al., 2013].

Peak calling. A genome-wide profile of TF binding is created by extending *aligned* short reads to the original DNA fragment length defined during the preparation of the DNA library (usually 150-200 bp). The extended reads pile up symmetrically on the complementary chromosome strands near the true TF binding sites [Pepke et al., 2009] and are aggregated into one strong single peak of the ChIP-Seq signal which identifies the potential location of the transcription factor binding site [Kharchenko et al., 2008, Leleu et al., 2010]. Benchmarks of peak calling programs [Szalkowski and Schmid, 2011, Wilbanks and Facciotti, 2010] recommended the MACS

program [Feng et al., 2012, Zhang et al., 2008] as the best peak caller for the identification of narrow TF binding sites.

The relationships between the strength of the ChIP-Seq signal and biological activity are currently being actively investigated [Zhang et al., 2012]. Some very active transcription factors display only modest ChIP-Seq signals. With such variation it is difficult to establish an *a priori* threshold for the number of peaks or the ChIP-Seq signal strength that would ensure the inclusion of all peaks containing the binding sites in a way that maximizes true positives and minimizes false negatives [Landt et al., 2012]. Some prominent peaks observed in ChIP-Seq data represent anomalies [Hoffman et al., 2013, Tewari et al., 2012]. Anomalous peaks are observed in the regions of open chromatin in all cell lines for most DNA binding proteins. They do not contain transcription factor binding sites and are accounted for by control data.

Motif Analysis. A motif is a DNA sequence pattern recognized by a protein. Motif patterns are modeled by position weight matrices (PWM) which contain an *a priori* set probability of each nucleotide in the motif sequence. PWMs quantitatively describe motif families characterizing the DNA binding sites of proteins with specific functions [Altobelli, 2012, Wasserman and Sandelin, 2004]. The presence of TF DNA-sequence motifs in peaks increases confidence that those peaks contain true TF binding sites [Lombard et al., 2011]. The peak regions can be scanned for the presence of known motifs recorded in TRANSFAC [Wingender et al., 1996], JASPAR [Portales-Casamar et al., 2010], STAMP [Mahony and Benos, 2007] and proprietary databases. Motif analysis tools search for the presence of known or *de novo* motifs by finding the DNA sequences in peaks that match the PWM model of the motifs [Aerts, 2012]. Motif identification tools relying on the PWM model use an expectation maximization algorithm to identify the motifs enriched under peak sequences. Tools based on inductive logic programming also describe the order in which found motifs tend to occur in peaks [Korol and Turcotte, 2012].

Preferred Distances in Composite Motifs. Genomic binding sites are found in specific

arrangements of transcription factor binding site clusters [Kulakovskiy et al., 2011], generally called Cis Regulatory Modules (CRMs) which can be considered to be composite elements. Transcription factors in protein complexes may have preferred binding distances with respect to each other [Kulakovskiy and Makeev, 2013]. For example, direct inspections of preferred distances in Spi-1 binding sites have shown dramatic positional preferences with respect to one other [Ridinger-Saison et al., 2012]. Different binding affinities of transcription factors have been associated with altered regulatory functions of transcription factors in different biological contexts and cell lines [Palii et al., 2011]. Composite motifs separated by a specific number of bases are longer and more specific than individual motifs. It has been argued that many identified TF binding sites are non-functional [Wasserman and Sandelin, 2004]. Longer composite sequences with specific spacing between the motifs, prevalent in peaks, provide more evidence that the peaks contain functional binding sites. For this reason it is very important to have tools that allow for the examination of composite sequences in peaks. No publicly-available tool with this functionality has been reported in literature.

Transcription Factor Regulated Genes. Transcription factor proteins bind to specific DNA sequences, recruit cofactors and modifiers to direct transcription initiation rates and control the expression levels of their target regulated genes [Aerts, 2012, Altobelli, 2012, Van Loo and Marynen, 2009]. Genome-wide TF binding sites obtained in ChIP-Seq experiments can be at various distances with respect to the gene promoter regions. The genes that have proximal TF binding sites are likely to be regulated by those transcription factors or their complexes through the recruitment of RNA polymerase to the sites [Struhl, 1999]. Target genes of TFs can be identified experimentally by inactivating the transcription factors and then measuring gene expression via microarray [Aerts, 2012]. The target genes can also be predicted computationally by examining which of the genes in the genome have TF binding sites in a close upstream or downstream proximity. The most useful tools to perform such computations are BedTools

[Quinlan and Hall, 2010] and the Bioconductor GenomicRanges package [Morgan et al., 2009]. Functions of the target genes can be characterized by using the Genomic Regions Enrichment and Annotation Tool (GREAT) [McLean et al., 2010]. Alternatively, the transcription factors associated with those genes can be examined with the ENCODE ChIP-Seq Significance Tool. This tool uses the database of transcription factors collected by the ENCODE project from 843 ChIP-Seq experiments across 90 cell lines [Auerbach et al., 2013].

1.1.2 Available Computational Pipelines with Integrated Computational Procedures to Analyze ChIP-Seq Data

The previously-discussed bioinformatics tools for ChIP-Seq data analysis are implemented in a variety of different languages and platforms. This poses considerable challenges to users such as biologists and clinicians who lack training in bioinformatics. Workflow management systems such as Galaxy [Blankenberg et al., 2011] and Taverna [Wolstencroft et al., 2013] offer partial solutions to these challenges by providing integrated environments hosting a variety of bioinformatics tools. Standard analysis workflows and data are shared by the community through these environments, which ensures reproducibility of analyses and results [Nekrutenko and Taylor, 2012]. The public servers Nebula [Boeva et al., 2012] and Cistrome [Liu et al., 2011] provide almost complete ChIP-Seq data analysis functionality through the Galaxy front end web interface. Another computational environment with already installed bioinformatics tools is available as a Linux machine image [Krampis et al., 2012]. Finally, several software packages for the integrative analysis of ChIP-Seq data are available as collections of separate modules.

The ChIPSeeker [Giannopoulou and Elemento, 2011] computational framework enables a broad analysis of ChIP-Seq peaks including the characterization of peaks with respect to CpG islands. It also provides comparisons with other published ChIP-Seq data sets, user-defined motif analysis in regulatory elements and comparisons of several different ChIP-Seq datasets.

The framework consists of command line tools, some of which can be accessed through the created user graphical interface.

Modules of ChIP-Seq data analysis are also offered by the Bioconductor packages PICS, rGADEM and motIV [Mercier et al., 2011]. In order to use all those tools a user must have high proficiency in R. The PICS package identifies peaks through the modeling of ChIP-Seq signal mixtures and excludes the regions in which short reads can't be uniquely mapped. The rGADEM package performs *de novo* motif searches. It uses a genetic algorithm to create position weight matrices of target motifs without requiring specification of motif size, and identifies enriched motifs with an expectation maximization algorithm. The motIV package describes identified motifs using data in STAMP [Mahony and Benos, 2007] and TRANSFAC [Wingender et al., 1996] databases. The ChIPpeakAnno [Zhu et al., 2010a] package provides means to analyze peaks with respect to genes and to retrieve gene-level functional annotations. The Sole-Search tool offers similar functionality through the web interface [Blahnik et al., 2010].

This section concludes the review of state-of-the-art methods and tools that are currently used in the analysis of ChIP-Seq data. Most frequently used software tools were summarized briefly in Table 1.2. The algorithms and their underlying methodologies were briefly discussed. The tools in the discussed pipelines provide the means to perform the majority of ChIP-Seq data analysis tasks. The importance of the identification of composite sequences in predicted transcription factor binding sites was highlighted. A shortage of algorithms and software tools to compute distances at which transcription factors prefer to co-occur was identified. By applying the steps of the discussed pipeline to the sequenced ChIP-Seq data we identify a list of genes that are targets of the studied transcription factor. Methods for the analysis of gene lists and gene-phenotype relationships are the subject of section 1.2.

1.2 Analysis of Gene Lists with the Enrichment of Gene Annotations

Large lists of genes are produced by high-throughput experiments aimed at identifying genes that are strongly related to or causative of biological phenomena under study. By what means can one start identifying those genes? The first step is to collect what is already known by looking up specialized information resources.

The Gene Ontology database is a major resource used by enrichment tools. It contains three categories of gene annotations: Molecular Function, Biological Process and a Cellular Component organized into a hierarchy [du Plessis et al., 2011]. Exhaustive information on gene annotations in various contexts (functional, pathway, interactions) is available through the portals of the National Center of Biotechnology Information (NCBI) [Maglott et al., 2011] and through many other resources of genomic information that are interconnected [Baxevanis, 2011]. Methods of *the analysis of the enrichment of annotations* of genes in large lists leverage information stored in those databases. Biological contexts that are most prominently associated with the gene list are discovered by computing the enrichment of gene annotations.

The underlying idea of enrichment of annotations analysis is that gene lists generated by high-throughput technologies contain **groups of genes with common features** involved in various biological processes. These common features are reflected by their annotations. Large numbers of genes in the list mapped by enrichment tools to the same or similar functional annotation serves as supporting evidence that those genes co-function in the same biological process or pathway.

Enrichment tools are characterized by the back-end database, the method to compute enrichment and presentation of the results. 68 developed enrichment tools for the systematic analysis of large gene lists were classified into Singular Enrichment Analysis (SEA), Gene Set Enrichment

Analysis (GSEA) and Modular Enrichment Analysis (MEA) [Huang et al., 2009]. Frequently used tools (as of September 5, 2013 in GoogleScholar) are listed in Table 1.3.

An individual list of genes is compared to a background by SEA tools. Usually genes in the

TABLE 1.3: Most Used Tools in the Enrichment Analysis of Large Gene Lists

Tools	**Citations**	**References**
SEA		
DAVID	4214	[Dennis et al., 2003]
EASE	1580	[Hosack et al., 2003]
GoMiner	1036	[Zeeberg et al., 2003]
BiNGO	992	[Maere et al., 2005]
GOStat	880	[Falcon and Gentleman, 2007]
FATIGO	883	[Al-Shahrour et al., 2004]
MAPP Finder	860	[Doniger et al., 2003]
GSEA		
GSEA	5027	[Subramanian et al., 2005]
PAGE	347	[Kim and Volsky, 2005]
GeneTrail	165	[Backes et al., 2007]
T-profiler	161	[Boorsma et al., 2005]
MEA		
topGO	409	[Alexa et al., 2006]
GENECODIS	187	[Carmona-Saez et al., 2007]
GOSim	96	[Frohlich et al., 2007]
ProfCom	41	[Antonov et al., 2008]
Enrichr	3	[Chen et al., 2013]

list are mapped to their GO annotations. Then, by applying a statistical test for each annotation, a number of genes in the list sharing that annotation is compared to a number of genes in the background also sharing that annotation. The test estimates the probability of observing that many genes in the list sharing the same annotation purely by chance. Enriched Gene Ontology annotations indicate biological themes linked to the investigated gene list. The Hypergeometric or Chi-Square or Kolmogorov-Smirnov tests are usually used.

The tools in the GSEA category analyze genes in a list as a group with respect to an ordered background list. A representative in this category is GSEA [Subramanian et al., 2005]. It is based on scoring the distribution of the ranks of the genes in the ordered background list.

The tools in the MEA category at first form groups of genes sharing similar annotations or related in some other way. For example, the genes in list that interact with each other are

represented as a module which is analyzed with SEA methods. The SerbGO website [Mosquera and Sanchez-Pla, 2008] provides useful guidance in selecting the most suitable tool for analysing the enrichment of annotations.

Analysis of the enrichment of annotations identifies biological contexts related to gene lists. Significantly large numbers of genes sharing the same annotations in a list as compared to a background serve as evidence that these annotations characterize collective function of the genes in the list. Such analysis can be regarded as unsupervised. Its results depend on information in back-end databases and chosen background genes. To infer relationships between genes and a phenotype of interest which is not covered by a back-end database, other methods, known as gene prioritization, are used. These methods are introduced in the following sections.

1.3 Gene Prioritization

Gene prioritization is a process of identifying the most promising gene candidates related to a phenotype of interest from a large list [Tranchevent et al., 2011]. A goal of the prioritization algorithm is to derive the connection between genes and a biological process of interest and rank the genes according to the strength of that connection. The top ranking genes are the suggested best candidates. Finding meaningful relationships between genes in a **large list** and a phenotype by manually reviewing the literature and genomic databases is very laborious and time-consuming. Efforts to automate this process have mostly been directed towards the prioritization of human disease genes [Masoudi-Nejad et al., 2012, Piro and Di Cunto, 2012] and less for model organisms and general phenotypes [Aerts et al., 2009]. Model organisms (including cells) provide experimental platforms for the investigation of developmental and cellular processes. This raises a need for algorithms that allow us to prioritize genes with respect to cellular phenotypes in biological contexts other than diseases [Yu et al., 2008a].

There are at least three dozen gene prioritization tools developed and maintained by various

research groups [Capriotti et al., 2012, Piro and Di Cunto, 2012, Tranchevent et al., 2011]. Gene prioritization tools differ from each other with respect to input and output data, the type of computational algorithms and the data sources used in the prioritization [Tranchevent et al., 2011]. Tools accept as input both free text descriptions of phenotypes and phenotype descriptions by genes with known functions. Candidate genes are prioritized either in the user-provided list or with respect to a whole genome. The tools use literature, protein-interaction networks, GO annotations, homology between the organisms, chromosomal regions and gene expression information as evidence to connect genes to phenotypes.

At a fundamental level, features that characterize a gene prioritization algorithm are: (**i**) how a phenotype is defined and (**ii**) how an algorithm identifies a link between genes and phenotypes. The definition of a phenotype will determine the rules by which the algorithm will mine available data resources to retrieve gene-phenotype links. In the following subsection, ways to define a phenotype are discussed in more depth.

1.3.1 Current State of Phenotypic Resources for Phenotype Definitions

The definition of a phenotype widely accepted in biology is *"the observable trait or the collection of traits of an organism resulting from the interaction of the genetic makeup of the organism and the environment"* [Mahner and Kary, 1997]. A phenotype means different things depending on the context and level of analysis [Marian, 2012]. In medical contexts the phenotype often refers to deviation from the *normal* [**?**]. In cellular contexts measurable cellular phenotypes are represented by features of cells such as the morphology (shape, size), the behavior (motility, growth), the developmental stage, the expression of specific genes and the rate of biochemical reactions [Carpenter and Sabatini, 2004, Fuchs et al., 2010]. Phenotype definitions have to be in some standard form in order to be used in computational applications.

Standardization of Phenotype Descriptions. Efforts to standardize phenotype descriptions [Bodenreider and Burgun, 2010, Houle et al., 2010] result in vocabularies of phenotypes. PhenomicDB database hosts textual phenotype descriptions collected after genetic screening [Houle et al., 2010]. Specialized limited phenotype vocabularies are available for model organisms [Flybase, 2013, Groth et al., 2010] and life sciences in general [Smith et al., 2007]. The GEN2PHEN knowledge base hosts phenotype information linked to genetic variation in humans and other model organisms in genome wide association (GWA) studies [Butte and Kohane, 2006, Webb et al., 2011]. Existing phenotype ontologies have limited use in gene prioritization so far [Bodenreider and Burgun, 2010], except for the Ontology of Human Diseases [Kohler et al., 2009]. Table 1.4 lists vocabularies that can be used as resources for phenotype definitions. In automated gene prioritization the definition of a phenotype should be formalized such that it could be used in computations. Two approaches to phenotype definition exist in gene prioritization. One consists of vocabularies of the domain-specific phenotype terminology usually implemented as ontology containing the concepts, the relationships between the concepts and the definitions of both [Gkoutos et al., 2005, Hoehndorf et al., 2012]. Another approach relies on a subset of genes with known functions. Currently, the Medical Subject Headings (MeSH) vocabulary created at the National Library of Medicine (NLM) [Bodenreider and Burgun, 2010, Groth et al., 2010, Huang et al., 2011] and the Mammalian Phenotype Ontology (MPO) [Smith and Eppig, 2012] are the best established controlled vocabularies encompassing curated concepts that can be used as phenotype descriptions.

TABLE 1.4: Controlled Vocabularies of Terms for the Phenotype Description

Ontology	**References**
Genome Phenome Network	[Houle et al., 2010]
Gene Ontology	[Ashburner et al., 2000]
Mammalian Phenotype Ontology	[Smith and Eppig, 2012]
Human Disease Ontology	[Kohler et al., 2009]
Medical Subject Headings	[Huang et al., 2011]
Open Biomedical Ontologies	[Smith et al., 2007]

Suitability of Medical Subject Headings Vocabulary to Define Phenotype. Phenotypic descriptions should be meaningfully associated with the biological processes that they denote. This is the main reason why MeSH terms are very well suited to serve as phenotypic descriptions. MeSH terms are assigned to the articles in the MEDLINE database by human curators to adequately reflect the content of each article. The MeSH vocabulary has approximately 23,856 descriptors (changing as the vocabulary is revised) divided into 16 categories. It has a hierarchical structure of levels in which broader terms encompass narrower terms as illustrated in Table 1.5 by a few examples.

MeSH terms are used in the Mouse Genome Informatics (MGI) database [Bult et al., 2013] which is a resource of gene-phenotype associations for the mouse organism. Genes in the MGI

TABLE 1.5: Categories and Examples of Broader → Narrower Hierarchy in MeSH Vocabulary

Category	**Examples of Broader → Narrower**
C. Diseases	Neoplasms → Tumor Virus Infections → Avian Leukosis Muskoskeletal Diseases → Muscular Diseases → Fybromyalgia
G. Phenomena and Processes	Cell Physiological Phenomena → Cell Physiological Processes → Cell Fusion Physiological Phenomena → Physiological Processes → Homeostasis → Body Temperature Regulation

database have been associated with phenotypes through the experimental manipulation of mouse genes. The phenotypes in the MGI are characterized by MPO terms and are mapped to the appropriate MeSH headings [Bult et al., 2013, Eppig et al., 2012, Smith and Eppig, 2012]. For example, *abnormal homeostasis* in the MPO maps to **Homeostasis** in MeSH, *decreased bone mineral density* in the MPO maps to **Bone Density** in MeSH.

To date the best characterized phenotypic domain is human diseases [Hamosh et al., 2005, Kohler et al., 2012a]. There is a shortage of phenotypic resources in domains other than diseases. Exceptions are the mouse [Eppig et al., 2012], drosophila [Flybase, 2013] and worm [Wang and Marcotte, 2010]. This shortcoming can be compensated for by using resources such as MeSH

vocabulary or by defining phenotypes according to subsets of genes known to be strongly related to those phenotypes. Known genes for phenotype definition are widely used in gene prioritization algorithms [Tranchevent et al., 2008, 2011]. However, if the phenotype of interest hasn't been well studied and does not have genes linked to it, then it is difficult or even impossible to use this approach. The way by which a phenotype is defined also determines the means the prioritization algorithm will use to relate genes to the phenotype.

1.3.2 Approaches that are Used by Gene Prioritization Algorithms

Domain-specific databases and literature are cross-linked with each other to infer relationships between genes and phenotypes [Frijters et al., 2010, Krallinger et al., 2010]. Text and data mining [Yu et al., 2008a], network analysis [Valentini et al., 2014] and genomic data integration [Aerts et al., 2006] are applied in this inference. Essential ideas underlying these inferences are based either on conceptual frameworks of linked data [Tiffin et al., 2009] or strategies of supervised machine learning [Moreau and Tranchevent, 2012].

An example of the linked data approach is Arrowsmith, in which concepts, co-occurring at high frequency in two disparate sets of literature articles, indicates meaningful links [Smalheiser and Swanson, 1998, Smalheiser et al., 2009]. The link suggests that fish oil can reduce Raynaud's syndrome symptoms, which has now been confirmed experimentally [Swanson, 1986]. An inference leading to this result was "fish oil reduces blood viscosity, platelet aggregations and vascular re-activity which are increased in Raynaud's syndrome" [Shatkay and Craven, 2012]. Prioritization algorithms, based on linked data, relate genes to phenotypes by using concept co-occurrences in literature and controlled vocabularies [Hristovski et al., 2005, Perez-Iratxeta et al., 2002a].

Supervised machine learning approaches are most successful in prioritization algorithms that define a phenotype through genes known to be related to the phenotype [Nitsch et al., 2011,

Tranchevent et al., 2008]. By applying machine learning (kernel methods are most used) a model characteristic of known genes is derived [Aerts et al., 2006, Tranchevent, 2011]. Genes are classified as similar or dissimilar compared to the model. Similar genes are called candidates and are ordered according to the degree of similarity [Aerts et al., 2006, De Bie et al., 2007]. The most used prioritization tools based on the GoogleScholar citations (as of September 2013) that are actively maintained are listed in Table 1.6. Three general categories can be distinguished: methods using training genes to define phenotypes, methods using keywords and methods working only with disease phenotypes.

TABLE 1.6: Tools Frequently Used in Gene Prioritization

Tools	**Web Link**
Define phenotype by training genes	
Endeavour [Aerts et al., 2006, Tranchevent, 2011, Tranchevent et al., 2008]	`http://www.esat.kuleuven.be/endeavour`
ToppGene [Chen et al., 2007, 2009]	`http://toppgene.cchmc.org/`
Define phenotype by keyword	
PolySearch [Cheng et al., 2008]	`http://wishart.biology.ualberta.ca/polysearch/index.htm`
CANDID [Hutz et al., 2008]	`https://dsgweb.wustl.edu/hutz/candid.html`
PosMed [Yoshida et al., 2009]	`http://omicspace.riken.jp`
Work with diseases phenotypes	
G2D [Perez-Iratxeta et al., 2002a, 2005]	`http://www.ogic.ca/projects/g2d_2`
GeneWanderer [Kohler et al., 2008]	`http://compbio.charite.de/genewanderer`
MimMiner [van Driel et al., 2006]	`http://www.cmbi.ru.nl/MimMiner/cgi-bin/main.pl`
SUSPECTS [Adie et al., 2006]	`http://www.genetics.med.ed.ac.uk/suspects/`
PhenoPred [Radivojac et al., 2008]	`http://www.phenopred.org`
GeneProspector [Yu et al., 2008b]	`http://www.hugenavigator.net/HuGENavigator/geneProspectorStartPage.do`

1.3.2.1 Methods Defining the Phenotype by Training Genes

Algorithms that define a phenotype through genes known to be related to the phenotype in prioritization use some similarity metrics between training genes and candidate genes. A representative in this category is Endeavour. Endeavour creates models from training genes using

different information sources. It compares descriptions of candidate genes in each information source with the model derived from training data and ranks the candidates for each data source individually according to computed similarity metrics. Individual ranks of candidate genes are transformed into rank ratios and for each gene a single rank value is derived via order statistics [Tranchevent, 2011]. All tools and their descriptions are presented in Table 1.7.

TABLE 1.7: Prioritization Tools Defining Phenotype by Training Genes

Endeavour
In Endeavour the phenotype is defined by training genes. It builds a phenotype model using different sources of genomic information derived from the training genes. Endeavour data sources consist of gene annotations, gene sequences, expressed sequence tags over multiple conditions, protein-protein interaction data and known transcription factor binding sites. The program works with the genes of human, mouse, rat, fly and worm organisms. It builds the model of the phenotype using information about the training genes in each of the genomic sources. It ranks the candidate genes according to how well they compare with the built model or by using the one class Support Vector Machine. Individual rankings in Endeavour are combined by order statistics [Tranchevent, 2011]. In the table of the ranked genes explanations are provided about the genes.
ToppGene
Candidate gene prioritization is one of the functions provided by the ToppGene tool. The user submits a set of training genes and a set of test genes. ToppGene first finds the significantly enriched annotations for the training genes in multiple data sources: GO annotations, literature, Interaction, Pathway, human and mouse phenotype data,TF binding sites, Cytobands, Coexpression Atlas, Drugs, microRNA and more. The candidate genes are ranked by the similarity of their functional annotations to the enriched annotations in the training genes. The similarity is computed as fuzzy-based measure [Popescu et al., 2006] or Pearson correlation coefficient. The user can examine the genes and the enriched terms of the the training set.
GeneWanderer
In GeneWanderer the candidate genes are retrieved from the genomic region given the genomic coordinates. The phenotype is defined either by the disease keyword or by the list of the training genes known to be related to the phenotype. If the phenotype is defined by the keyword then the known genes associated with it are retrieved. The tool measures the distance between the candidate genes and the training genes in the protein-protein interaction network. The tool is specific to the human diseases.
SUSPECTS
The algorithms in the SUSPECTS is based on the assumption that genes involved in the complex phenotype will have similar domains, annotations, patterns and expressions. The candidate genes can be defined by chromosomal coordinates or gene names. The phenotype is defined either by the list of known genes or a disease keyword. In the later case the SUSPECTS will form appropriate training gene data set by retrieving relevant genes from OMIM, Human Gene Mutation Database and Genetic Association Database. The SUSPECTS tool ranks the candidate genes by the similarity of their GO annotations, Interpro protein domains and expression profiles to those of the training genes. The ranking score is computed as a weighted average of scores from all sources.

1.3.2.2 Methods Defining the Phenotype by Keywords

Algorithms describing phenotypes by keywords in prioritization use frequencies of gene-associated documents that have keyword matches. Documents are retrieved from literature or the local databases of the tool. Some tools in this category also provide the user with the option to define phenotype through training genes. The number of matching documents is a very imprecise indicator of the strength of a link between a phenotype and a gene since the list of keywords does not define the phenotype in a very meaningful way compared to definitions in the terms of controlled vocabularies. All tools and their descriptions are summarized in Table 1.8.

1.3.2.3 Methods Prioritizing Genes for Disease Phenotypes

Algorithms in this category are designed to prioritize genes with respect to disease phenotypes. A representative in this category is G2D. It uses OMIM identifiers for phenotype description and maps identifiers to corresponding MeSH terms associated with diseases. G2D infers the links between MeSH terms to establish a chain of evidence connecting disease phenotypes to GO annotations that identify the best candidate genes in given genomic regions associated to human diseases. Tools and their descriptions are summarized in Table 1.9.

1.3.3 Inference of Relationships between Genes and Phenotypes

Prioritization methods produce ranked list of genes that are the best candidates in relation to the phenotype of interest. Rankings are based on evidence scores computed by a prioritization algorithm for each candidate gene. Only a few algorithms (PosMed, ToppGene, G2D) allow comprehensive examination of evidence underlying the inferred gene-phenotype relationship. In PosMed the evidence is provided by the content of documents associated to gene-phenotype relationships. ToppGene lists the properties of training genes. In G2D the inference can be explained as illustrated in Figure 1.3 through the real example of exploring candidate genes associated with Cleft Lip disease.

TABLE 1.8: Gene Prioritization Tools Defining Phenotype by Keywords

PolySearch

PolySearch allows queries in the form of: *Given X find all Y*. X and Y can be diseases, tissues, cell compartments, gene/protein names, SNPs, mutations, drugs and metabolites. If the phenotype is defined by keywords, then PolySearch retrieves the documents matching all keywords in the Pubmed, OMIM, DrugBank, Swiss-Prot, Human Mutation Database, Genetic Association Database and Human Protein Database. A ranked list of requested biomedical entities that are associated with the text of the query is returned. The score of the entity is proportional to the number of document matches in the databases. Users can browse the results and examine the matching publications and sentences.

CANDID

The CANDID tool ranks the whole human genome. The phenotype can be defined as a free text, for example `Myoblast Cell Fusion`. CANDID retrieves all articles from PubMed matching all words in the submitted text. Gene association with the textual phenotype description is performed by using NCBI gene2pubmed database looking for the matching text in publication metadata associated with genes. The genes are ranked according to the number of associated publications. Publications associated with mouse genes homologous to the human genes are also included in the ranking. Other data sources that can be used to rank genes are protein-protein interaction, conservation information, gene expression in tissues and SNP associations. The output is a list of the ranked human genes with scores computed for each selected data source. Examination of the evidence is not supported.

PosMed

Positional Medline is the semantic engine that ranks biomedical entities by the statistical significance of the associations with the provided keywords. The strength of the associations between biomedical entities and keywords is based on the number of the documents they share. The document categories comprise MEDLINE (9378134 documents, content: MEDLINE titles, abstracts and MeSH terms), REACTOME (3809 documents, content: Pathway information from REACTOME), Protein-protein interaction (73645 documents, content: Protein-Protein Interactions in Human and Mouse from IntAct and Arabidopsis from AtPID), Gene ontology (12787 documents), Human disease ontology (2282 documents), Mammalian phenotype ontology (7440 documents), Microarray based co-expression data for Arabidopsis. Given the keyword defining the phenotype and the type of biomedical entity to score (either gene or metabolite or drug) the PosMed returns a list of the scored entities linked to the phenotype, sorted according to the strength of the connection between them. The PosMed supports human, mouse, rat, arabidopsis and rice organisms. The user can browse through all documents of the established links.

The association between Cleft Lip and the rs987525 variant from region 8q24.21 has been replicated independently in several different populations [Nikopensius et al., 2009] but no associated gene was found. The MYC gene was suggested as good candidate gene by G2D in that genomic region. The OMIM phenotype Cleft Lip was mapped to the corresponding MeSH C terms. The link between the gene and the disease was inferred through the relationship between the terms **Craniofacial abnormalities** and **Homeodomain proteins** and the relationship between the term later and the GO annotation a **Sequence specific DNA binding transcription factor**

TABLE 1.9: Gene Prioritization Tools for Disease Phenotypes

G2D
In G2D the disease phenotype is defined by the OMIM identifier which is mapped to the associated MeSH terms of the diseases. The candidate genes are selected from the provided genomic region possibly containing a marker associated with the disease phenotype. G2D establishes a chain of evidence connecting the disease phenotype to the genes by forming the links between the terms in MeSH and GO annotations. The MeSH terms of the disease (category C) are linked to the MeSH terms of the chemicals and drugs (category D) which are linked to the Gene Ontology annotations. The connections between the terms are established by computing the normalized frequency of MEDLINE documents in which the MeSH terms (C and D) occur together. The connection between the protein GO annotations and the MeSH D terms are established by computing the normalized frequency of co-occurrence of GO and MeSH D term in papers supporting experimental evidence for the GO annotation. The GO annotation receives a combined score which is used for ranking of the candidate genes. The flow of inference is shown in Figure 1.3. G2D outputs the list of the strongest candidates with all supporting evidence. G2D has an option to define the phenotype by the known genes.
MimMiner
MimMiner works specifically with OMIM. Genes are retrieved associated with the OMIM disease identifier. The user can examine the genes in the Uniprot Database.
PhenoPred
PhenoPred uses known gene-disease associations, protein-protein interaction data, protein functional annotation and protein sequence data as the data sources. The Human Disease Ontology is used to specify the disease names. The tool ranks the gene given the disease name, alternatively it can rank the diseases for the provided gene. Browsing the evidence is not supported.
GeneProspector
GeneProspector is a genetic epidemiology oriented web portal to perform searches for links between human diseases and genes. It works with a local disease database Phenopedia (2608 diseases), the gene database Genopedia (12413 genes), Literature (84821 articles), Investigator (205803 investigators), Genome wide association studies (9929 GWAS) Cancer (13420 records). A user can compute the risk of the disease given the genotype and to find the prevalence of the genotypes in populations. The Gene Evidence option provided the disease retrieves the linked genes in the priority order. The prioritization score is computed from the number of various types of the articles in genetic epidemiology literature in which disease and gene appeared together normalized by the total number of the articles in the database.

activity of the MYC gene. The MYC gene is regulated by the CTCF transcription factor [Qi et al., 2003] which has a binding site at the genomic location of rs987525 thus suggesting the possibility that this SNP marker might indeed be linked to the Cleft Lip phenotype through regulatory interaction with the MYC gene [Schaub et al., 2012]. This relationship is speculatory, it hasn't been experimentally confirmed.

Another connection between Cleft Lip OMIM phenotype and the BMP4 gene was identified through the relationship between the Cleft Lip and the term **Bone Morphogenetic Protein 4**

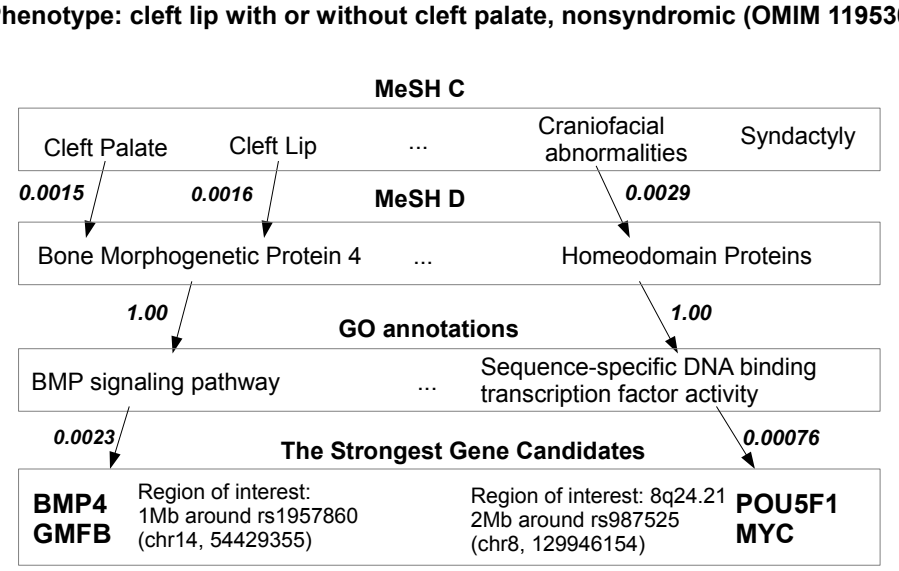


FIGURE 1.3: Connections Computed by G2D in Prioritization of Genes with Respect to the Cleft Lip Disease

which is related to the GO annotation **BMP signalling pathway**. The BMP4 gene harbors the rs1957860 marker variant which is known to be associated with Cleft Lip [Suzuki et al., 2009].

Inferences of gene-phenotype links in G2D are based on the computation of fuzzy binary relationships between the MeSH terms from the MEDLINE and NCBI Entrez databases and GO annotations [Perez-Iratxeta et al., 2002a, 2005] (Mathematical details underlying fuzzy binary relationships will be discussed in Section 3.2 of this thesis). At first, the OMIM disease phenotype in G2D is mapped to the corresponding MeSH terms of Disease category C. Then, meaningful links are formed between these MeSH C terms and MeSH D terms that denote molecular entities - drugs and chemicals. These links are discovered by analyzing the frequencies at which MeSH C and MeSH D terms co-occur together as annotations of articles in the MEDLINE database. Gene functions provided by GO annotations are also meaningfully related to MeSH D terms. This is done by analyzing of how frequently genes, annotated by particular GO terms, are associated with the articles that are annotated by particular MeSH D terms in MEDLINE.

Gene-phenotype relationships which are established by G2D through intermediate links such as **phenotype to chemical to GO annotation to gene** may also be used in explaining the nature of those relationships.

1.3.4 Concluding Remarks

This overview of gene prioritization methods revealed two streams in phenotype definition: methods that use training genes to define phenotypes and methods that define phenotypes by keywords. Methods, defining phenotypes by training genes, have utilized well established machine learning techniques and proved to be successful in several applications as described in references [Aerts et al., 2006, Tranchevent, 2011]. However, there are many cases in which no genes have been associated to phenotypes. Therefore, prioritization algorithms relying on training genes are not applicable. However, methods that use keywords for phenotype definitions are underdeveloped. In these methods the number of keyword-matching documents associated to genes usually serves as a criterion of the strength of association which is a rather superficial estimate of relationships.

Currently existing phenotypic resources such as ontologies and controlled vocabularies of phenotypic terms are well developed only for the domain of human diseases [Kohler et al., 2009, 2012b]. Active development of gene prioritization tools is also directed towards the prioritization of genes with respect to human diseases [Moreau and Tranchevent, 2012, Piro and Di Cunto, 2012]. Tools that would allow users to prioritize genes in a variety of organisms with respect to more general phenotypes or in domains of cellular processes are not developed.

Gene prioritization algorithms produce lists of the best candidates which are most strongly related to the phenotype of interest according to the criteria set by the algorithm. These results have greater practical value if the algorithm provides means to examine the evidence leading to them. The multipurpose systems PolySearch and PosMed have means to examine resulting

outputs. However, the formed links between concepts are not as comprehensive as one would expect. G2D is the most comprehensive in this respect.

A concluding result of this overview is that there is a shortage of tools to prioritize genes with respect to a general phenotype that can't be defined by known genes. Suppose a researcher has a large list of experimentally obtained genes which is related to a cellular phenotype such as myoblast cell fusion that can't be defined by known genes. Then he/she would not be able to use any of the existing tools to perform a meaningful prioritization of his/her gene list, except to do of a general analysis of the enrichment of annotations.

1.4 Objectives, Contributions and Outline of this Thesis

The work presented in this thesis addresses two specific shortcomings in the area of bioinformatics algorithms. It is organized into two parts, each corresponding to a major objective.

Objective 1. To create a gene prioritization algorithm. Analysis of state-of-the-art approaches in gene prioritization identified a shortage of algorithms to prioritize genes with respect to non-disease phenotypes defined by meaningful keywords (see Section 1.3.4). This thesis shows how an algorithm was created and implemented to address this shortage, so that gene prioritization can be done in situations where little is known about the phenotype and it has no known related genes. The following aims were formulated to achieve Objective 1:

- To define a phenotype by a meaningful keyword. To do this, MeSH terms denoting biological processes were extracted from the NCBI database and prepared for use in the algorithm.

- To develop a strategy of linking genes to phenotypes. To achieve this, the algorithmic principles of fuzzy set theory underlying G2D were used to relate genes within Gene Ontology terms to phenotypes described in MeSH terms. Semantic relationships between

these terms were modeled on fuzzy binary relations (FBR) **inclusion** such that broader phenotype terms include narrower terms denoting gene function. This is a main novel contribution to the field of bioinformatics algorithms by this thesis as FBR of inclusion are for the first time applied to link concepts used to describe phenotype and gene function.

- To implement an algorithm in which phenotype can be described by known genes. To achieve this, the Gene Ontology database was extracted and reformatted for use in the algorithm.

- To develop a web application for the completed gene prioritization algorithm to allow users to examine the evidence supporting the resulting prioritization. A back-end database and dynamic web page were created to achieve this aim.

Objective 2. To implement and expand a ChIP-Seq data analysis pipeline by creating a novel tool to analyze the distances between transcription factors in a composite motif. Analysis of state-of-the-art approaches to ChIP-Seq data analysis revealed that existing pipelines do not provide means of flexibly and comprehensively analysing the spacing between transcription factor motifs (See Section 1.1.2 and paragraph **Preferred Distances in Composite Motifs** in Section 1.1.1). It is important to be able to compare the tendency of transcription factors to co-occur in peaks at preferred distances against the way they bind in background sequences. This option is currently not available in existing ChIP-Seq data analysis tools. The following aims were formulated to achieve Objective 2:

- To create and implement a suite of algorithms for the analysis of composite sequences in peaks. To achieve this, new and unique tools to analyze the spacing between transcription factors in peaks were created in R - a novel contribution of this thesis to the field of bioinformatics algorithms. The tools were validated via application to two ChIP-Seq data

sets: unpublished BCL11B transcription factor data and published transcription factor TAL1 data.

- To create a computational analysis pipeline for a ChIP-Seq experiment and integrate it into the Galaxy workflow management system to ensure reproducible results. To achieve this, some open-source programs were chosen out of a well-researched group of candidates for creating the pipeline. Then, the functions missing from their software libraries were programmed. Finally, the command line tools of the pipeline were integrated into Galaxy.

The algorithms developed in this thesis can be used separately or together. The end goal of the analysis pipeline is to identify lists of genes that may be targeted for regulation by transcription factors. These lists can be analyzed further with the unique gene prioritization algorithm introduced here. Both the gene prioritization tool and the newly completed ChIP-Seq pipeline are available as downloads in tarball form.

This thesis is organized as follows. Chapters 2 and 3 deal with methods. In Chapter 2, the optimized ChIP-Seq data analysis pipeline is described. In Chapter 3, the gene prioritization algorithm, the underlying theory and the benchmark created to test the algorithm are discussed in detail. In Chapter 4, a practical application of the pipeline and gene prioritization algorithm, in the form of an analysis of ChIP-Seq data from the BCL11B and TAL1 transcription factors, is illustrated. Chapter 5 presents general discussion and concludes the thesis. Technical details and computer scripts are included as appendices.

Chapter 2

Optimized ChIP-Sequencing Data Analysis Pipeline

This chapter addresses implementation of pipeline for the analysis of data produced in ChIP-Seq experiments aiming at the genome-wide characterization of transcription factors and their target genes. Our pipeline was used to analyze data sets produced by the Illumina HiSeq2000 sequencing platform with samples containing up to 100 million sequenced reads.

Public servers hosting computational tools for ChIP-Seq data analysis [Blankenberg et al., 2011, Boeva et al., 2012, Liu et al., 2011] lack integrated tools for the analysis of composite motif sequences and functions to perform annotations of genomic regions in terms of their co-location with features such as promoters, untranslated regions, exons and introns, and the association of genomic regions to genes within varying distances. This deficit has been addressed by implementing new tools for the analysis of composite sequences and integrating required and publicly available stand-alone programs through the Galaxy interface.

The pipeline elements for quality control, alignment, peak calling and the visualization of aligned reads in the UCSC genome browser are implemented as single programs called through the command line interface. Tools for the analysis of genomic regions were written in the R and

Perl programming languages and were integrated into the local Galaxy platform. This integration supported the exchange of intermediate data files between tools and the reproducibility of the analysis.

The choice to divide the pipeline into command line tools and tools integrated into the workflow management system was motivated by the long processing times needed to perform alignment and peak-calling on large ChIP-Seq data sets. The first part consists of command line shell scripts that can be executed in parallel using high-performance computing infrastructure if available. The second part - the downstream analysis of transcription factor binding sites - is more efficient through the workflow management system. This choice was motivated by the need to run the same analysis several times with different parameters and by the need to reuse and share intermediate data. All of the pipeline tools can be used as single programs or applied in batches to data processing.

2.1 Materials and Methods. Elements of ChIP-Seq Data Processing Pipeline.

The developed ChIP-Seq data analysis pipeline will be described in the following sections as a two part approach. In the first part, the command line tools will be reviewed. Command line tools are used to perform computationally intensive tasks. Some tools can take advantage of available high performance computer cluster architecture and can process data simultaneously on several computer cores. In the second part, the tools developed for genomic region analysis will be reviewed. These tools were integrated into the Galaxy workflow management system. The user interfaces and parameters of these tools are explained in Appendix B. The dependencies of the developed pipeline are summarized in Section B.10.

2.1.1 Command Line Tools

The tasks in the computationally intensive parts are summarized in Table 2.1.

TABLE 2.1: The Elements of the ChIP-Seq Data Analysis Pipeline as Command Line Tools

Tasks	**Subtasks**
Pre-processing of the raw ChIP-Seq data	Remove the reads with poor quality bases Remove the reads contaminated by linkers and adapters Trim the short reads of poor quality segments
Mapping the reads to the reference genome by MAQ or BWA programs	Split the data into multiple files containing 2 to 4 million reads Align sequentially or submit the alignment jobs to execute in parallel Merge the aligned files and convert them into binary BAM format Transform the aligned reads into wig format for visualization Visualize the aligned reads in the UCSC genome browser
Genome wide identification of the transcription factor binding sites: peak calling	Choose peak calling program (default MACS) Identify peaks Assess sufficiency of the reads to cover the identified peaks

2.1.1.1 Preprocessing Task

The raw data acquired from the sequencing equipment must be of proper quality for further processing. The quality of the raw data is characterized by (i) the number of reads containing poor quality base calls, (ii) the number of sequences contaminated by linkers and adapters, (iii) the number of duplicated sequences. Reads containing many base calls of poor quality should be removed because such reads may align to the wrong positions in the reference genome. The reads contaminated by linkers and adapters in the library preparation step should be either removed (default in this work) or trimmed, because they put unnecessary computational burden on alignment tools [Koboldt et al., 2010, Leleu et al., 2010]. In trimming, the known sequence

of the adapter or linker is removed from the start or the end of the read. Reads trimmed to less than 26 bp in length should be discarded because short reads may have many different matches in the reference genome; this would introduce ambiguous mappings.

Short reads from massively parallel sequencing instruments can be identical in their sequences but differ in their base-call qualities. The identical sequences may contain high quantities of the contaminated reads as well as the original experimental reads. There are two possibilities. Either one discards the identical reads before the alignment step or after the alignment step. The rationale for discarding the duplicated reads before the alignment step is to reduce computational burden and the time needed for alignment. The argument in favor of aligning the duplicates is that one is not losing any potential information since the reads having variable base quality might have originated from different genome locations. However, the second strategy is prone to result in ambiguous mapping. For this thesis, the duplicated reads were discarded before alignment. Identical sequences with identical base-qualities are collapsed to a single read.

The implemented programs process sequencing files that have the FASTQ format [Cock et al., 2010]. The FASTQ format consists of four lines: a short read identifier, the sequence content, an empty line and a base-calls quality line. The quality control step in our pipeline consists of simple Perl scripts to process the FASTQ files and conditionally remove bad quality reads. Listing B.19 in Appendix B illustrates the code for conditional removal of poor quality reads. The removal of contaminated and duplicated reads follows a similar scheme with different conditional statements.

2.1.1.2 Task of Alignment to a Reference Genome

To obtain a profile of transcription factor binding sites, the reads must be mapped onto the reference genome. Typically, short read data sets contain millions of reads of fixed length from 26bp to 150bp. The data processed by our pipeline contained 50bp reads. For ChIP-Seq data it

is important to align the short reads accurately. Both MAQ [Li et al., 2008] and BWA programs [Li and Durbin, 2009, 2010] are accurate alignment programs. The MAQ program was shown to be robust against a higher number of mismatches that may originate from sequencing errors [Hatem et al., 2013]. The BWA works with paired end sequencing data. It is one of the most used alignment tools and its use is well documented [Zhou et al., 2013].

In our pipeline we used the MAQ and BWA alignment programs. A computer cluster with the Sun Grid Engine (SGE) queue management system was used to execute alignment tasks in parallel [Oracle, 2010]. Listing B.20 in Appendix B shows how to prepare to run an alignment task by MAQ in parallel. In order to submit a file for alignment with MAQ it should be in Sanger FASTQ format [Cock et al., 2010]. The files containing the short reads and the reference genome should be transformed into binary MAQ format. MAQ works best with 2 million reads; therefore, a large file should be split into smaller parts. Listing B.21 in Appendix B illustrates a code snippet to submit an alignment task for parallel processing. The aligned short reads represent only part of the original fragment of DNA. To visualize the alignment in the genome browser the reads have to be extended to their original length and transformed into GenomeGraph representation. The GenomeGraph format merely represents the density of short read tags along the chromosomes in the genome. Preceding the extension and transformation the obtained alignments should be converted into BED format [Quinlan and Hall, 2010]. Listing B.22 in Appendix B shows how to prepare an alignment file for visualization in the UCSC genome browser.

2.1.1.3 Task of Peak Calling

The short reads of the aligned ChIP-Seq data produce peaks of varying heights at certain genome locations across the reference genome. The short reads pile up symmetrically on complementary chromosome strands near the true transcription factor binding sites [Kharchenko et al., 2008, Pepke et al., 2009] and produce a dense strong signal as in Figure 1.2. The reads on the

complementary strands are shifted halfway from the centers of the pileups on the complementary strands to the middle of the interval connecting them. They are aggregated into one strong, single peak of a ChIP-Seq signal at the specific genomic location, which identifies the true location of the transcription factor binding site [Kharchenko et al., 2008, Leleu et al., 2010]. The transcription factor binding profile consists of the background regions and the peak regions that contain potential transcription factor binding sites.

Identifying true transcription factor binding sites is driven by the nature of the ChIP-Seq experiment. The proteins with narrow binding sites have very different binding profiles compared to those of chromatin modifying proteins. The aligned ChIP-Seq data is processed by special software designed to automatically identify these sites. Some tools such as PeakRanger [Feng et al., 2011b] and PeakSeq [Rozowsky et al., 2009] are better suited for the identification of broad peaks. Some tools such as MACS [Zhang et al., 2008] or SISSRs [Jothi et al., 2008] are recognized as more suitable to predict sharp narrow/point binding sites of transcription factors [Laajala et al., 2009, Leleu et al., 2010, Wilbanks and Facciotti, 2010]. In the later versions of MACS, the modes to process both narrow and broad peaks [Feng et al., 2011a] have been implemented.

The genomic regions encompassing open chromatin sites are enriched for all transcription factors and in all cell lines [Tewari et al., 2012]. A ChIP-Seq control performed with a nonspecific antibody [Liu et al., 2010a] shows high enrichment in those regions. Therefore, those regions represent false positives and should be discarded. Peak calling algorithms distinguish those regions from predicted true TF binding sites, and then discard them, by comparing the reads enrichment of the signal over the control.

We used the MACS (ver. 1.3.7) tool in our pipeline as a peak caller. The algorithm in MACS builds a peak model versus a background model from empirical ChIP-Seq data. It employs a sliding window to adapt to varying ChIP-Seq signal strengths. It takes into account the

distances between peaks on complementary strands to determine the correct transcription factor binding site location [Kharchenko et al., 2008, Zhang et al., 2008]. In each sliding window MACS determines the distance between the peak summits mapped to the opposite DNA strands. Then, it shifts the reads to aggregate the read tags from both of the strands. The middle point of the shifted interval represents the peak summit. Secondly, MACS fits a Poisson distribution in order to model the original peak tag distribution at nucleotide positions in the peak region. The peak model is built in the first pass. In the second pass the program identifies the genomic locations which are likely to be peaks, comparing them to the fitted model of the peak. Each region identified as a peak is assigned a p value estimating its significance. As for parameters, MACS uses the DNA fragment length used for computing a shift and a predefined fold of tag enrichment over the background. Depending on the parameters MACS may identify a large number of peaks. False positive calls can be identified as having larger p values and low tag densities in the peaks [Landt et al., 2012].

MACS has parameters that affect predictions. The most important parameters are displayed in Table 2.2. Depending on the parameters the MACS program may identify a large number of peaks. The $mfold$ parameter controls the stringency of the model in terms of the enrichment magnitude of the signal over the background. In some experiments the true signal may not be strong [Zhang et al., 2012] and the $mfold$ has to be adjusted to produce a reasonable peak model for the particular ChIP-Seq data. The window size controls the width of the region used to build the peak model. The window size is determined by the original fragment length of the sonicated DNA. Usually the DNA fragment length is identified experimentally in the DNA library preparation step. If the DNA fragment length is not available it is reasonably approximated by the average length of the DNA fragments in the library. In a recent study a method was developed to estimate DNA fragment length by correlating the enriched ChIP-Seq signals accumulated on opposite DNA strands [Ramachandran et al., 2013]. The commands to

TABLE 2.2: Parameters in the MACS Program Controlling the Prediction of Peaks Containing True TF Binding Sites

Parameter	**Description**
Window size **bw**	This parameter controls the size of the sliding window of 2*bw=2*(DNA sonication fragment length) size in which the program identifies regions enriched by the mfold over the number of tags in background.
Shift size **d**	This parameter by default is estimated by MACS to identify the distance (bp) between the peak summits originating from the read alignment on the opposite DNA strands. MACS then shifts the reads on both strands by $d/2$ towards the $3'$ ends to estimate the region of the TF binding site.
Fold of enrichment **mfold**	This parameter sets in advance the magnitude of enrichment in the signal versus the background required to estimate the parameters of the peak model. The peak model is subsequently used to compare against the regions of the entire ChIP-Seq signal and identify the regions that are significantly similar to it. The default is 32.
Peak significance **pvalue**	The significance of the identified peak based on a comparison of the enriched raw ChIP-Seq peak to the model. The default value in the MACS program is p value=1e-5.
DNA short read length **tsize**	This parameter identifies the original short read size.

call the MACS program are shown in Listing B.23 in Appendix B.

2.1.1.4 Estimation of Sufficiency of Coverage

Some ChIP-Seq experiments produce poor data but it is too costly to repeat the experiments. Thus, an assessment about whether there are sufficient reads to identify the transcription factor binding profile can be performed. In order to estimate the sufficiency of the mapped reads for the identification of TFBS, an approach described in [Palii et al., 2011] was included in the proposed pipeline. The assessment is carried out by computing the rate of conversion of mapped reads from the background into the signal as new reads are added.

Reads that occur in peaks are designated as a signal. The reads that do not occur in peaks are designated as a background. All available reads in the aligned ChIP-Seq data set are divided into

n equal parts. In the first iteration of the algorithm the peaks are identified using the first 1/n of the data. Reads from this data fraction that occur within the peaks are considered the signal. In subsequent iterations the remaining subsets of the data are added one by one to the read set and peak identification is carried out. In each iteration the fraction of reads in peaks is considered the signal. This process is repeated until all of the data has been added. In each iteration the ratio of signal reads to the number of reads used for peak calling is computed as the rate of the background conversion into the signal. If the rate levels out as more and more reads are added to the initial data set then this indicates that no information is gained by increasing the amount of data for peak calling. This means that the available reads sufficiently cover the genomic features which we aimed to identify. Such a test provides information of whether the available data is sufficient to find genomic features of interest. The program implementing coverage computation is listed in Appendix B Listing B.24. The application of this tool is illustrated in Chapter 4. The plot of the rate of the background-to-signal conversion for BCL11B peaks is shown in Figure 4.2.

2.1.2 Integrated Tools

This part describes the tools that were implemented using different programming environments and integrated into the workflow management system. The tasks that are performed using these tools are summarized in Table 2.3.

2.1.2.1 Characterization the Genome-Wide Binding of the Transcription Factor

The identified genome-wide peaks containing potential TF binding sites are located in different genomic segments. An assessment of how many peaks are found in gene promoters, exons, introns and intergenic areas provide insight about the possible way in which the TF carries out regulatory functions. Such characterization also increases our confidence as to whether the identified peaks contain realistic TFBS. It is expected that a considerable fraction of the peaks will be found

TABLE 2.3: The Elements of the ChIP-Seq Data Analysis Pipeline that Were Integrated into the Galaxy Environment

Tasks	**Subtasks**
Analysis of the peaks	Summarize the genomic features (promoters, untranslated regions, exons, introns) occurring in peak regions. The newly integrated tools: • `Closest Bed`, • `Peak Genomic Features`, • `Peak counts over Genomic Features`. Associate the peaks with the genes. The newly integrated tool: • `Peak gene association according to the given criterion`. Analyze lists of associated genes Compute the distribution of peak distances from the Transcription Start Sites of the closest gene
Analysis of the motifs in identified peaks	Summarize all occurrences of the given short sequence of interest in all peaks. The new integrated tool: • `seq_in_peaks`. Identify *de novo* the TF sequence motifs in peaks. The newly integrated tools: • `Fasta from Bed`, • `ishuffle`, • `motifs_denovo`. Identify composite sequences of interest present in peaks Explore the preferred binding distances of the two TF motifs of interest in the peaks and background. Integrated new tool: • `pref_dist_in_peaks`.
Analysis of the derived gene lists. Future work	Perform Enrichment of Annotations Analysis Perform Gene Prioritization with respect to the phenotype of interest

within or close to gene promoters; in general in eukaryote genomes the promoters are key elements regulating gene expression [Struhl, 1999]. The genomic features that characterize the peaks are promoters, untranslated regions, exons and intron. The peak may span over several genomic features, only one of which is associated with the peak. The genomic features are associated

with peaks with the following precedence [Palii et al., 2011]: promoter, 3'untranslated region, 5'untranslated region, exon, intron and intergenic area.

To assign the genomic features to the peaks and compute the summary, three programs were created and integrated into Galaxy: `Closest Bed`, `Peak Genomic Features` and `Peak Counts over Genomic Features`. The genomic coordinates of all documented gene accessions are extracted from the RefSeq UCSC gene table. Then, for each peak the closest or intersecting gene regions are identified. Each genomic feature intersecting the peak is identified and recorded. The peak region can intersect multiple genomic features: the promoter, the 5' and 3' untranslated regions, the exons, the introns and the intergenic area. However each peak is associated only with one genomic feature. Genomic features are associated with a peak based on the precedence: promoter, 3' UTR, 5' UTR, exon, intron and inter-genic area. For example, if a peak intersects an inter-genic region, and a promoter, and a 5' UTR, then it will be associated with the promoter. If a peak intersects an exon and intron, then it will be associated with only the exon. User interface and tool outputs are explained in Appendix B.11.

2.1.2.2 Task of Association of Peaks to Genes

Transcription factors binding to gene promoters regulate the level of expression of those genes. Transcription factors bind to promoters, introns and distant intergenic regions and perform their regulatory functions directly or through recruitment of co-factors. The promoter region is commonly regarded to be within 5Kb upstream of the gene transcription start site (TSS) [McLean et al., 2010]. However, this distance is arbitrary. Promoters can also be found in wider areas with respect to the gene TSS, including regions downstream of the gene 3' end. To identify the genes that are possible regulatory targets of a transcription factor the association of peaks to genes must be explored at various distances. The GREAT tool [McLean et al., 2010] offers gene-peak association functionality and allows changes to rules of gene-peak association. The

gene-peak association in GREAT is a part of this tool which makes it challenging to use if one needs to perform many gene-peak associations with different rules.

Our pipeline includes a tool to associate genes with peaks. Given a list of the genomic regions of peaks (in BED format) and the desired distance criteria, the tool computes the regions around each gene called query regions. The genes for which the query regions intersect the peaks are associated with those peaks. The distance can be defined related to the TSS or the full length of the gene. Specific query regions of various sizes can be used to associate genes with peaks. This is useful if there is a need to associate genes with peak regions containing transcription factor binding sites that are distant from the genes. The interface and output of the tool are explained in Appendix B.12.

2.1.2.3 Task of Motif Analysis in Peaks

Motif analysis is an integral part of the characterization of a transcription factor. Specific motifs present in peaks can also indicate the actual in vivo binding of a particular protein [Lombard et al., 2011]. There are two types of motif analysis with respect to peak regions: exact motif matching and *de novo* motif discovery. *de novo* motif discovery finds over-represented sequence patterns in sequences of peak regions. Motif matching detects occurrences of specified sequences in peaks. The motifs can be represented as (i) a position frequency matrix, (ii) a position probability matrix, (iii) a position weight matrix and (iv) a position specific scoring matrix [Wasserman and Sandelin, 2004]. All motif models are based on the frequency of the occurrence of a particular nucleotide at a specific position in the sequence. The motif analysis tools in this pipeline use a position frequency matrix (PFM) of motifs to characterize the motif and are implemented in R. The Bioconductor motifRG [Fong et al., 2012] package was chosen for *de novo* analysis of the motifs since it has a convenient interface to call motif analysis functions and is implemented in R. The utility of motifRG in TFBS identification was demonstrated in several applications

[MacQuarrie et al., 2013, Sebastian et al., 2013, Yao et al., 2013].

The motif tools in the current pipeline are used to perform four tasks: extraction of sequences in FASTA format for given genomic regions (the `Fasta from Bed` tool), randomization of DNA sequences (the `ishuffle` tool), finding all occurrences of a particular pattern in peak sequences (the `seq_in_peaks` tool) and *de novo* pattern search in peaks by motifRG (the `motifs_denovo` tool). The tools use motifs specified by consensus sequence using IUPAC code [Morgan et al., 2009].

The `Fasta from Bed` tool extracts DNA sequences from provided genomic regions. The `ishuffle` tool randomizes the provided genomic sequences. This tool preserves the original frequencies of dinucleotides. The `ishuffle` tool is used to create background sets for the motif discovery tool `motifs_denovo`. The tool `seq_in_peaks` finds all the occurrences of a motif pattern given as a consensus sequence in the DNA sequences of the genomic regions. The search is performed using the forward and reverse versions of the pattern. This tool can be used to summarize the variants of the consensus sequence pattern in peaks. The *de novo* motif identification tool is a wrapper of the motifRg [Fong et al., 2012] Bioconductor package. The tool has a lengthy execution time for large numbers of motifs (10 and above). The inputs and outputs and tool interfaces are described in detail in Appendix B.13.

2.1.2.4 Analysis of Transcription Factor Composite Sequences

In different biological contexts the same transcription factors may be characterized by different binding affinities. For example, it is known that in Jurkat cells the EBOX motif (5'-CANNTG-3') and GATA motif (5'-WGATAA-3') prefer to co-localize at 7,8 and 9 base pairs apart [Palii et al., 2011]. Whether transcription factors have tendencies to preferentially bind at certain distances must be explored using computational analysis preceding the intended in vitro testing. Such computational analysis consists of matching the composite motifs to the peak sequences

identified in the ChIP-Seq experiment. Existing preferred binding distances will manifest in a majority of composite motifs matching sequences of the peaks by comparing them to the background, in which a preference to bind at a specific distance cannot be higher than expected by chance. The most informative background in such a case is the whole genome; since it is not biased by a subset of regions, a whole genome binding profile of composite motifs can be obtained. The `Preferred Distance` tool was implemented in the pipeline because at the time of the thesis work there was no convenient publicly available tool to perform analyses of composite motifs and the distances between them.

The `pref_dist_in_peaks` tool was implemented in R. It creates a list of composite motifs and performs an exact pattern matching of them to the DNA FASTA sequences of the peaks and in chromosome 1. The chromosome masked for repeats is used as background in the `pref_dist_in_peaks` tool. There are four types of identified matches. The peak sequences are tested in the original and complementary forms. The composite motifs are matched to those sequences in both the forward and reverse orientations. All matches of the composite motifs found in the peak DNA sequences are recorded and used to compute the frequencies of the observed distances between the motifs. Relative motif orientation is accounted for by the user providing the required sequence orientation in specification of the anchor and other motifs. The frequencies at which the composite with the particular distance between the motifs occur are normalized by the total number of matches. A plot which displays the frequencies of the occurrences of the composite motifs at distances upstream [-20..0]bp and downstream [0..20]bp of the anchor motif in the peaks and Chromosome1 is created. An example of a plot of preferred distances is in Figure 4.4 in Section 4.1. The interface to the tool, its input and output are described in Appendix B.18.

2.2 Conclusions

The developed pipeline tools will assist in characterizing the binding of a studied TF genome-wide. The most time-consuming computational part is alignment to the reference genome. The alignment was prepared to run on a computing cluster in parallel using the Sun Grid Engine environment. The tools to perform associations of genes to peaks and to analyze the occurrences of motifs in peaks were integrated into the Galaxy environment, which provides an easy-to-use interface. Many simple repetitive analyses can be performed quickly and easily through this interface. The expedited results of the tools can be reused by other tools making the analysis even more efficient. Because of the modular environment ensured by the Galaxy environment, existing tools can be improved or changed without affecting the functionality of the whole system or their integration with other tools.

Chapter 3

An Algorithm for the Prioritization of Candidate Genes

The developed gene prioritization algorithm aids the analysis of how candidate genes may relate to a phenotype of interest in situations where little is known about it. This chapter describes the components, underlying theory, validation and applications of this algorithm.

First, the overall structure of the algorithm is presented and its components described in the Materials and Methods Section 3.1. The developed algorithm consists of two components. The first component is based on analysis of annotations of articles in MEDLINE literature. It uses phenotype definitions developed from MeSH keywords and the MEDLINE literature database as data sources for the inference of links between genes and phenotypes. The second component is based on training genes. It uses certain training genes known to be related to certain phenotypes to define a phenotype. In this component the gene-phenotype links are inferred by estimating the extent that the GO annotations of known genes and annotations of genes in a genome of a selected organism are semantically similar. Both of these annotation categories produce a weighted list of GO annotations. The weights are computed indicators of how strong the association is between the functional annotations of genes and the phenotype of interest. As a result,

annotated genes can be ranked in order of the weights of their annotations; the candidate genes will be ordered in terms of strongest to the weakest with respect to the phenotype of interest.

Second, the theory based on fuzzy set theory underlying the algorithm,which works by inferring relationships between genes and phenotypes, is outlined in Section 3.2. In order to contain the theoretical concepts related to this, a background is provided on the utilization of fuzzy binary relationships in retrieving contextual information from document collections in Subsection 3.2.2. This background is important because it conveys the nature of a semantic relationship modeled by fuzzy binary relationships. The following subsections 3.2.3 and 3.2.3.2 present a mathematical formalization of the computations of gene-phenotype relationships and illustrate this computation with real examples.

Third, the validation strategy and process to ensure the performance of the developed algorithm are described in Section 3.3. The algorithm was validated using gene sets from the Mouse Genome Informatics database, known to be related to the chosen phenotype through knock-out experimental evidence. The validation was carried out separately for both components of the algorithm - the literature-based component and the component based on training genes.

Fourth, the results that were obtained are presented in Section 3.3.5.

Then, Section 3.4 illustrates the application of the developed gene prioritization algorithm with respect to the cell fusion phenotype. The algorithm was also used to prioritize a list of genes reported in literature as strongly related to the Breast and Colorectal cancers to illustrate its application to disease phenotypes. Lastly, the advantages, biases and limitations of the proposed algorithm are discussed in Section 3.5.

A comparison of the literature-based gene prioritization method and the CANDID prioritization tool is presented as a supplementary material in Section A.4 for human genome genes associated with diseases. Implementation details and and elements of a user interface are included as an Appendix.

3.1 Materials and Methods: Components of the Prioritization Algorithm.

The developed algorithm consists of several components and processing modules. A chart illustrating these components and the flow of information between them are shown in Figure 3.1. At the input, the user provides (i) a definition of a phenotype either by using a MeSH keyword

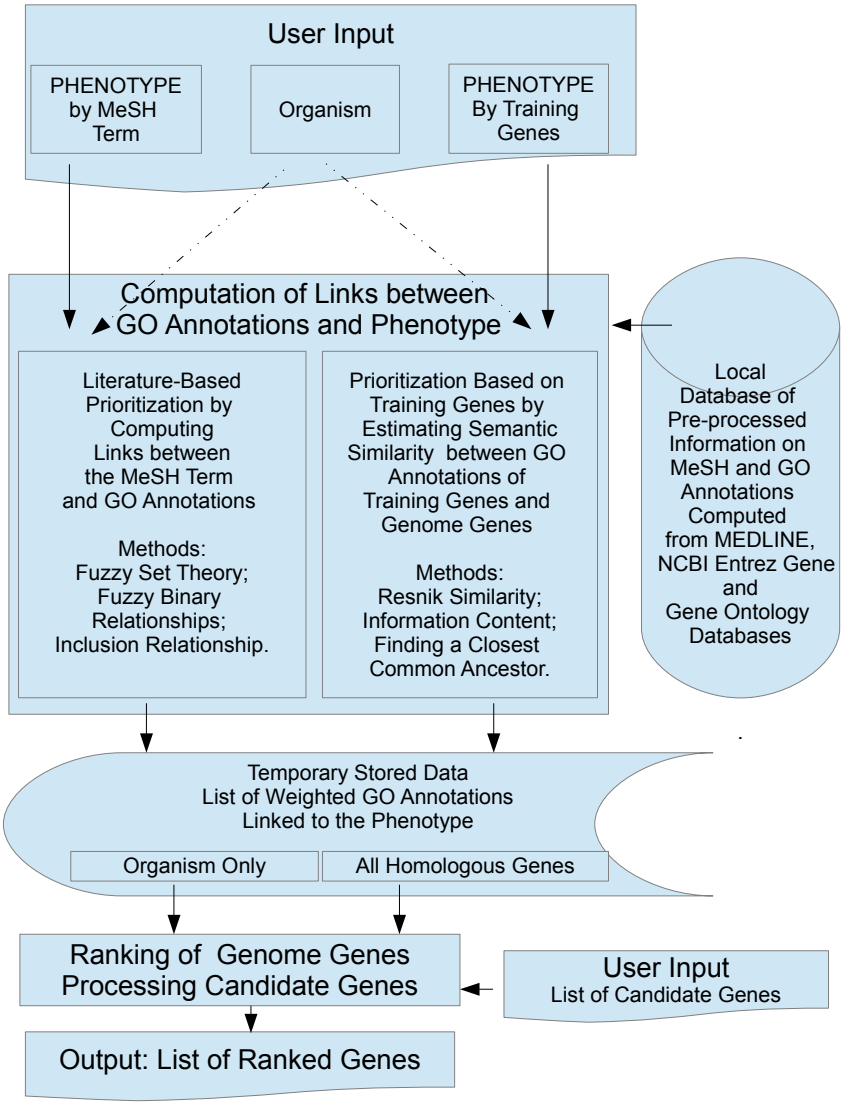


FIGURE 3.1: A Flowchart of the Developed Gene Prioritization Algorithm and the Relationships between its Components.

or with a list of training genes, or both, and (ii) an organism, the genes of which are being prioritized. The algorithm works with human, mouse and fly genomes. Optionally, the user can provide a list of genes to be ranked with respect to a phenotype of interest.

Given the phenotype and the genes the prioritization algorithm has to deduce the links between them by mining the information sources selected. In the proposed algorithm the main information sources are: a vocabulary of Medical Subject Header (MeSH) terms created by the National Library of Medicine (NLM) [Groth et al., 2010], MEDLINE literature and the NCBI Entrez Gene [Maglott et al., 2011] and Gene Ontology [du Plessis et al., 2011] databases. Depending on the inputs, either one or two processing modules are invoked to compute the links, using a local database storing pre-processed data about the co-occurrences of MeSH annotations denoting phenotypes and chemicals in MEDLINE articles. The database also stores information about the co-occurrences of GO and MeSH annotations from the NCBI Entrez Gene database as well as information about the ancestors and descendants of each GO annotation of the entire Gene Ontology for the human, mouse and fly. The result of this computational step is a list of weighted GO annotations that are found to have links to the phenotype throughout the entire Gene Ontology. The weights indicate the strength of the connections between each GO annotation and the phenotype denoted by the MeSH terms or training genes.

In the second stage of processing, the genes of the entire genome of the selected organism are ranked in order of the weights of their GO annotations. The more heavily-weighted GO annotations indicate strong connections between the annotation and the phenotype. The most heavily-weighted of all annotations assigned to that gene is used in ranking. If there is a tie, then the second largest weight is used. For each tie the same procedure is applied. If a list of candidate genes was provided, then these genes are ranked in accordance with their positions within the already-ordered genome. The genes at the top of the ranked list are the best candidates with respect to the phenotype of interest. The algorithmic principles underlying these

computations, their inputs and outputs are discussed in detail in the following subsections.

3.1.1 Description of Input to the Gene Prioritization Algorithm

In the prioritization algorithm, the user must provide the organism for which the prioritization is being performed, but the phenotype definition can be set by inputting either a keyword or a list of training genes known to be related to the phenotype of interest. For the former option the proposed algorithm uses Medical Subject Headings (MeSH) vocabulary which is organized hierarchically so that the broader terms span over more specific terms, as is illustrated in Table 1.5 in Chapter 1. The categories of MeSH terms that are used to define phenotypes in this algorithm are described in Table 3.1. The last row of the table shows how many terms are included in the current version of the algorithm. The terms in the Anatomy category designate

TABLE 3.1: MeSH Categories that are Used to Define a Phenotype

Phenotype Category	**A (Anatomy)**	**C (Diseases)**	**D (Drugs and Chemicals)**	**G (Processes and Phenomena)**
Type	cells, tissues, organs and bodily systems	diseases and abnormal processes	processes mediated by drugs and chemicals	explicitly defined chemical and biological processes
Example	Bone Marrow Cells	Leukemia	Central Nervous System Depressants	Cell Motility
Number of Terms	1668	4510	8958	1909

phenotypes as cells, tissues, organs and bodily systems. For example, one might be interested in which candidate genes are most strongly linked to the "**Bone Marrow Cells**". The phenotypes covered by the terms in the Drugs and Chemicals category may be used to indirectly describe phenotypes as responses to the effects of specific chemicals. For example, one may want to know which genes are strongly related to the "**Central Nervous System Depressants**". The algorithm uses $20,246$ GO annotations that characterize genes of the human, mouse and fly organisms in the NCBI Entrez Gene database.

Computing the relationships between phenotypes and functional gene annotations requires

data about co-occurrences of MeSH terms and GO annotations. These data were collected from the MEDLINE, NCBI Entrez Gene and Gene Ontology databases and stored locally. The database component of the prioritization system makes data management independent of a user interface and facilitates scaling of the system in the future. Tables of the local database were created by using NCBI *E-utilities* [Sayers, 2009] and custom scripts. The data sources and local database content (as of September 2013) employed by the prioritization algorithm are described in Appendix A.

3.1.2 Computation of Gene-Phenotype Relationships Based on Literature

The prioritization algorithm uses MeSH terms to define the phenotype in both the **non-disease** and **disease** domains. The main idea underlying the proposed algorithm is that a phenotype defined by the MeSH G term as a biological process or phenomenon is meaningfully related to a subset of the MeSH D terms denoting molecular entities such as drugs and chemicals.

Similarly, the gene functions encoded by GO annotations might be meaningfully related to molecular entities, denoted by MeSH D terms, through the chemical processes in which these gene functions might be affected. These relationships can be derived from MEDLINE.

The strength of the relationships between MeSH terms denoting phenotypes and MeSH D terms denoting molecular entities can be quantified by using the frequency with which these MeSH terms co-occur in MEDLINE articles. Similarly, the strength of the relationships between GO annotations and MeSH D terms, denoting molecular entities, can be quantified by using the frequency with which the MeSH D terms and GO annotations in genes associated with both co-occur. Whether a link between the MeSH D term and a gene exists can be determined from MEDLINE articles that are associated with that gene in NCBI *pubmed2gene* database.

Figure 3.2 presents an example of how links between the "**Cell Fusion**" phenotype and GO annotations are formed resulting in a list of ranked genes. The mathematical framework of this

approach used is outlined in Section 3.2. In this example, the "**Cell Fusion**" phenotype is related

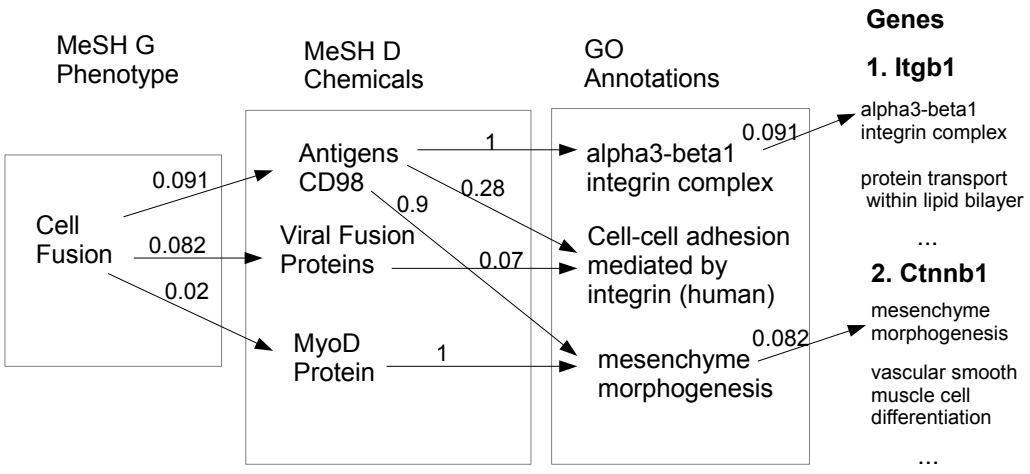


FIGURE 3.2: Example of Inferences in Literature-Based Gene Prioritization Algorithm Using MeSH Keywords to Define the Phenotype of Interest

to three molecular entities denoted by the MeSH D terms: "**Antigens CD98**", "**Viral Fusion Proteins**" and "**MyoD Protein**". The strongest relationship is with "**Antigens CD98**" and its value is 0.091. This phenotype is also related to a set of GO annotations of genes through these three intermediate connections.

Only one path exists that connects the "**Cell Fusion**" phenotype to the GO annotation "**alpha3-beta1 integrin complex**", which passes through "**Antigens CD98**". The combined values of the intermediate relationships 0.091×1 result in the weight of 0.091 assigned to the "**alpha3-beta1 integrin complex**" GO annotation.

Two paths connect the "**Cell Fusion**" phenotype to the GO annotation "**mesenchyme morphogenesis**". This annotation is assigned the weight of 0.082 which is computed via the intermediate relationship with "**Antigens CD98**". That value is larger than that of 0.02 computed via the intermediate relationship with "**MyoD protein**".

Two genes **Itgb1** and **Ctnnb1** in this example are annotated by the GO terms "**alpha3-beta1 integrin complex**" and "**mesenchyme morphogenesis**". The weights computed for these annotations - 0.091 and 0.082 - indicate the strength of the relationship between them and the "**Cell Fusion**" phenotype. The genes are ranked by the magnitude of the weights assigned to their annotations. The strongest candidate linked to the "**Cell Fusion**" phenotype in this example is the **Itgb1** gene which appears at the top of the ranked list. Relationships of other annotations to these genes with the phenotype of interest have not been computed. The formal computation for Figure 3.2 is outlined in Table 3.2.

TABLE 3.2: Computation of the Relationships between Phenotype and GO Annotations.

Formal Description of the Diagram in Figure 3.2

Suppose that there is a set of chemicals $d_j, \quad j = 1, \ldots, N$ through which the phenotype and GO annotation are linked. $V1_{gd_j}$ is the strength of the relationship between the MeSH term g of the phenotype and the MeSH D term d_j of the chemical. $V2_{d_j go}$ is the strength of the relationship between the MeSH D term d_j of the chemical and the GO annotation go. The final weight assigned to the GO annotation go is computed by multiplying the values of the strength of the relationships $V1_{gd_j}$ and $V2_{d_j go}$ for each pair connected through d_j and taking the maximum $w_{go} = \max_j(V1_{gd_j} \times V2_{d_j go})$. In this way the weight encoding the maximum strength of the relationship between the phenotype and the GO annotation is derived. The theory underlying the computation of $V1$ and $V2$ relationship values is covered in Section 3.2.

In computing the relationships between a given phenotype and GO annotations, the weights of all annotations in the Gene Ontology are evaluated. Optionally, the weighted GO annotations of the selected organism can be augmented with the annotations of homologous genes in other organisms. In this way, the sensitivity in detecting existing relationships between genes and phenotypes is increased. A putative example is as follows: no relationship exists between the intellectual disability phenotype and annotations of the tweek gene homolog in humans. But this relationship exists between that phenotype and annotations of the tweek gene in flies (*Drosophila Melanogaster*).

3.1.3 Computations of Gene-Phenotype Relationships Based on Training Genes

A phenotype can be defined by training genes known to be causing or strongly related to the phenotype of interest. The training genes on the whole should adequately represent the phenotype with their functional characteristics. This is important because the candidate genes are prioritized by estimating their functional similarity to the training genes. This is estimated through the semantic similarity between their GO annotations. Such estimation is possible because gene annotations in Gene Ontology form a hierarchy in which the annotations that have related meaning are also in close proximity in the hierarchical taxonomy.

The most general concepts such as "**biological process**" and "**molecular function**" and "**cellular component**" are at the root of the hierarchy. As we go from the top to the bottom of the hierarchy the terms become more and more specific. Figure 3.3 shows the taxonomy in Gene Ontology that contains the term "**endothelial cell migration**". The terms above it in the hierarchy are ancestors. The closest common ancestor of "**endothelial cell migration**" and the "**regulation of cell migration**" is "**cell migration**". The semantic similarity between two GO annotations is defined by their relative position in the taxonomic hierarchy. Each GO term go_i together with all its descendants annotates a certain number of genes n_i. The root node "**biological process**" and all its descendants annotate all N genes. For our algorithm we use the definition of Information Content [Kohler et al., 2009, Resnik, 1999] for a go_i annotation as:

$$IC(go_i) = -\log(\frac{n_i}{N}). \tag{3.1}$$

General terms at the top of the taxonomy close to "**biological regulation**" or "**locomotion**" have little information content since they, with all their descendants, annotate many genes. Very specific GO terms and their descendants annotate rather few genes and have high information content. For example, the term "**endothelial cell migration**" together with all its descendants

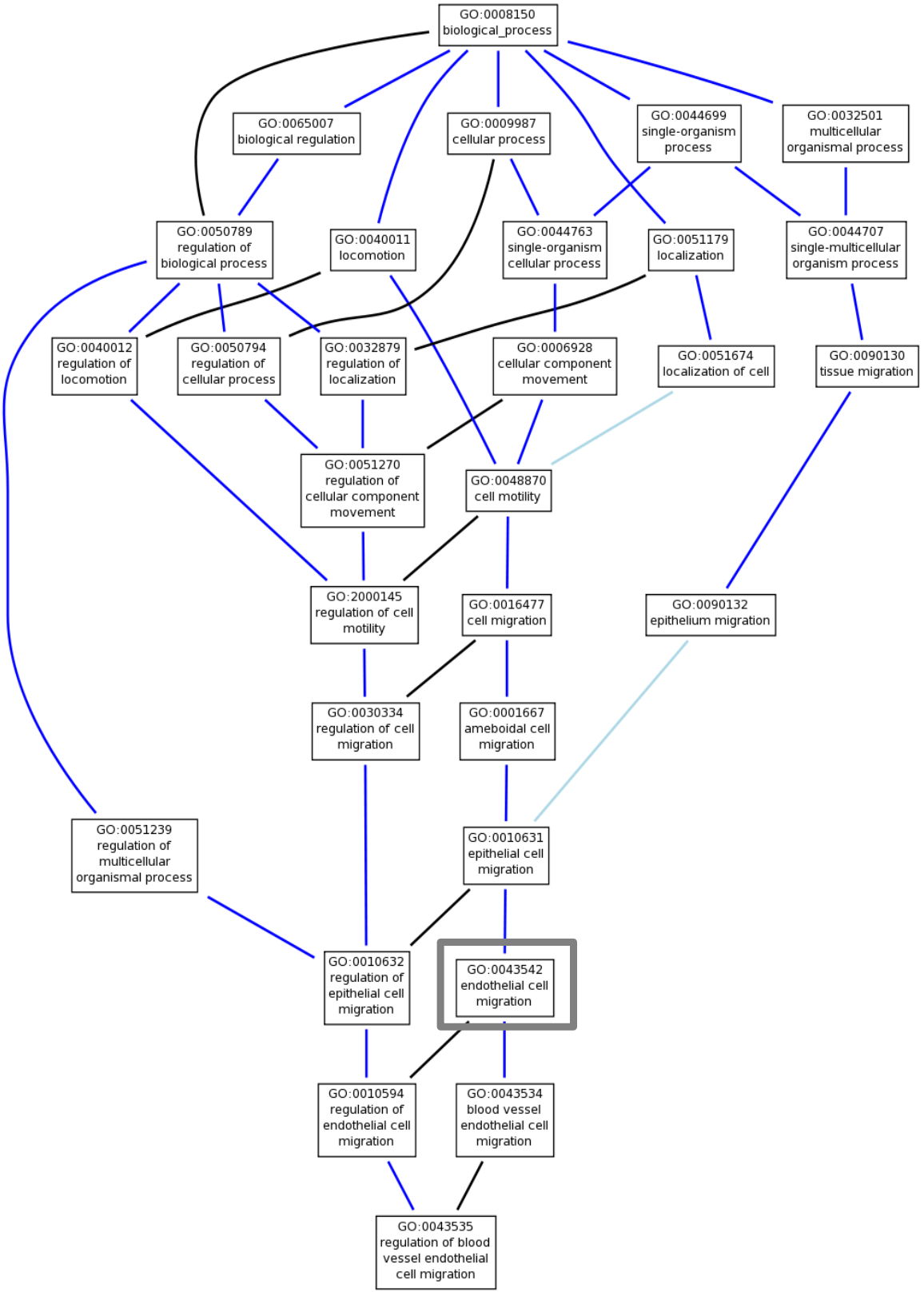


FIGURE 3.3: Subtree of Gene Ontology Containing "**endothelial cell migration**"

annotates 276 gene products in human, mouse and fly. Its information content is high (6.699).

By using hierarchical relationships between the terms in Gene Ontology we can estimate the semantic distance between two terms [Resnik, 1999]. This is defined as the information content of the most informative common ancestor [Houle et al., 2010, Kohler et al., 2009]. For example, the most informative common ancestor of the terms "**regulation of cell migration**" and "**endothelial cell migration**" is "**cell migration**" with information content:

$$IC(cell\ migration) = -\log(2731/223963) = 4.407. \tag{3.2}$$

In the proposed algorithm, the pooled GO annotations of the training genes are used to compute the semantic similarity between them and all other Gene Ontology annotations. The genome genes of human, mouse and fly are ranked by the semantic similarity scores of their annotations.

3.1.4 Ranking and Prioritization of Candidate Genes

Gene Ontology annotations describe a gene qualitatively in terms of its functional properties and cellular localization. In the proposed algorithm genes are linked to phenotypes through GO annotations by assigning numerical weights to them. The weights indicate the strength of the relationship between the phenotype and the GO annotation. Each algorithm computes the weights of GO annotations in a different way.

The algorithm component based on phenotype definition by training genes computes the weights by estimating the semantic similarity between the GO annotations of the training genes and of the entire Gene Ontology (see Section 3.1.3). The component based on phenotype definition by the MeSH keyword computes the weights from the links between the MeSH terms inferred by using MEDLINE (see Section 3.1.2). The same GO annotation will have different weights if assigned by different algorithms.

The relationship of a gene to the phenotype is quantified by a number of GO annotations

of that gene that were assigned the weights and also by the magnitude of those weights. Figure 3.4 presents a schematic of gene characterization by their weighted GO annotations. The genes are ordered according to the magnitude of the weights w of their annotations. In the proposed algorithm only the maximum weight is used in ranking. The rationale of this choice is explained in Section 3.2. If there is a tie, then the second largest weight of the annotations is compared.

In both parts of the prioritization algorithm (literature-based and based on training genes)

Genes	**Gene Ontology Annotations**			
Gene A	**GO 1** $w1 = 0.9$	**GO 2** $w2 = 0.77$		
Gene B	**GO 11** $w11 = 0.7$	**GO 12** $w12 = 0.5$	**GO 13** $w13 = 0.48$	**GO 14** $w14 = 0.2$
Gene C	**GO 101** $w101 = 0.55$	**GO 1** $w1 = 0.41$	**GO 103** $w103 = 0.15$	

FIGURE 3.4: Illustration of Principle of Gene Ranking by Using Weights Assigned to their GO Annotations. These Annotations are Assigned Weights by the Prioritization Algorithms.

the genes of the whole genome of the selected organism are ranked. The best candidates related to the phenotype of interest are found at the top of the ranked genome because their annotations have the largest weights and therefore the strongest relationship with the phenotype.

Each part of the algorithm may identify and rank a different number of genes. To make the gene rankings comparable across algorithms the ranks of the genes are transformed into a percentile of the distribution of ranks. Let N be the total number of genes in the ranked list. Let r be the rank assigned to the gene by the prioritization algorithm. The percentile is $pt = \frac{r}{N}$. The percentile representation is termed a relative score in G2D [Perez-Iratxeta et al., 2002a]. In other publications it is termed rank ratio [Bornigen et al., 2012].

The result of the prioritization algorithm is an ordered list of the genes in the whole genome according to the strength of their relationship to the phenotype of interest. If the user provided a

list of gene candidates to be prioritized, the genes will appear at specific positions in the ordered list of genome genes. These positions will most likely be different in ordered lists computed by different algorithms. In individual rankings the best candidates are genes which appear at the top positions in the ordered list. Individual rankings can be combined into a single score. In the proposed algorithm the combined score for the gene is a minimum relative score of *pt* out of all scores. Table 3.3 presents a hypothetical example of the prioritization of candidate genes with respect to a phenotype. In this example, gene A is the best candidate that has the smallest combined relative score. Gene A in Table 3.3 was placed on top ranking first by Algorithm

TABLE 3.3: A Hypothetical Example of the Prioritization of Candidate Genes. Algorithm 1 Ranked 10 Genes. Algorithm 2 Ranked only 8. The Rank of the Gene C is not Available (NA) for Algorithm 2. The Score Is a Percentile. The Combined Score Is the Smallest Percentile.

Gene	**Score(Rank) Algorithm 1**	**Score(Rank) Algorithm 2**	**Score Combined**
Gene A	0.1 (1)	0.625 (5)	0.1
Gene B	0.3 (3)	0.25 (2)	0.25
Gene C	0.4 (4)	NA	0.4
...	...	...	...
Total Genes	10	8	10

1, but only as fifth by Algorithm 2. Gene B was ranked as second from the top by Algorithm 2, but only as third by Algorithm 1. In this example Algorithm 2 identified and ranked only 8 genes. The rank for Gene C is not available for Algorithm 2. The score of each gene represents the relative score *pt* as previously defined. The combined score is the minimum *pt*. It represents the overall best attained score for the gene.

3.1.5 Availability of Gene Prioritization Algorithm

The link to the algorithm is available through figshare repository `http://figshare.com/phenotypelinks`. The user interface and particular implementation details are summarized in Appendix A. The

running time depends on the size of the training gene set and the number of links that the phenotype has with the MeSH D terms.

3.2 Mathematical Framework Underlying the Literature-Based Prioritization Method

MeSH terms and GO annotations are natural language concepts. In order to link a phenotype to a gene function we have to define a relationship between these concepts. The mathematical framework, allowing us to work with linguistic information and perform logical inferences on concepts, is based on fuzzy set theory [Zimmermann, 1996]. Fuzzy set theory, including computing with linguistic variables, attempts to formalize and model human reasoning in imprecise environments and has applications in control systems engineering, signal processing and information retrieval [Zadeh, 2001, Zimmermann, 2010].

Concepts of natural language can be represented by fuzzy sets. Such representations enable logical inferences on linguistic terms by applying mathematical operations on those fuzzy sets. In the proposed algorithm semantic gene-phenotype relationships are modeled by fuzzy membership functions. The links between the concepts denoting a phenotype and gene function are inferred by applying the fuzzy binary operations of the fuzzy sets defined on these concepts.

This section presents the theoretical framework and formalism on which the literature-based prioritization algorithm is built. The concepts of fuzzy set and modeling of semantic relationships between the concepts by fuzzy binary relationships are described in more depth in order to better explain a rationale for choosing this stronger approach in the developed gene prioritization algorithm.

3.2.1 The Fuzzy Set

The basic construct of the fuzzy set theory is a fuzzy set with its defining membership function. Let X represent a collection of objects x. Let A be a set defined on X. If the set A is crisp, then the object x either belongs to the set A or not. If it belongs to the set then its membership equals 1. The object's membership in A equals 0 if it does not belong to the set. Fuzzy set differs from the crisp set by allowing objects to have a partial membership in it. If set A is fuzzy, then the object x belongs to the set A to some extent. This extent is given by the membership function $m_A(x)$ defining that fuzzy set as $A = \{(x, m_A(x))|x \in X\}$.

The classical illustration of the concepts of the fuzzy set and the membership function is the representation of age by two fuzzy sets **young** and **old** depicted in Figure 3.5. The fuzzy

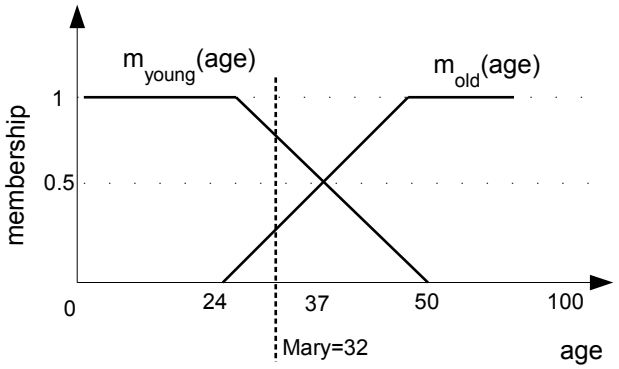


FIGURE 3.5: Fuzzy Sets **young** and **old** Representing the Age.

sets **young** and **old** of the people are defined by the membership functions $m_{young}(age)$ and $m_{old}(age)$. Each person, according to their age, belongs to both sets to some degree defined by the membership functions. A person of the age of five has membership 1 in the set **young** and 0 in the set **old**. Mary, according to her age, has a larger membership $m_{young}(32) = 0.75$ in the set **young** than in the set **old** $m_{old}(32) = 0.25$. The values of the membership functions can be interpreted as the truth values of the statements *Mary is young* and *Mary is old*.

Fuzzy sets consist of meaningfully grouped objects according to some proximity criterion, for example, a person's age. In another example, let X be a set of articles in MEDLINE. The articles in X about `homeostasis` can be defined as a fuzzy set H with some membership function. The membership function can be defined as the frequency of the word `homeostasis` in each article. Each article in X will belong to the fuzzy set H to a degree proportional to the frequency of the word `homeostasis` in the article. Articles that do not mention `homeostasis` will have zero membership in H.

3.2.2 Fuzzy Binary Relation

For information retrieval from large document collections, the content of those collections is structured by creating some metadata describing that content. This metadata can be associated with individual documents or with the entire document collections and is used for indexing. A thesaurus is an example of such metadata [Miyamoto, 1990, Perez-Iratxeta et al., 2002b].

The thesaurus is a dictionary in which a title keyword is described by associated descriptor keywords [Miyamoto, 1990]. The title and descriptor keywords can be related in several ways. The title keyword can be a broader or narrower term than the descriptors (inclusion) or the title can be somehow associated with the descriptors (similarity). The source data from which the descriptors are retrieved usually consists of collections of articles [Miyamoto, 1990, Perez-Iratxeta et al., 2002b]. The relationships between the descriptors emerge from the content of the articles in the collections. The article collections may cover different domains. The same words will have different title-descriptor relationships in different thesauri built from domain-specific document collections. Therefore, in a thesaurus meaningful relationships between keywords are automatically built in through title-descriptor representations.

For example, let us consider words in the phrase `magnetic field current resonance`. Let the word `magnetic` be the title keyword and let the other words be the descriptors. Each of

these words has an independent meaning. However, in a specific context, the meaning of the title keyword `magnetic` can be related to the meanings of the other descriptors `field`, `current` and `resonance` in a specific way, where the meaning of `magnetic` encompasses the meaning of other words. For example, in a collection of articles talking about general science the individual words `magnetic`, `field` and `resonance` may appear very frequently. However, the phrase `magnetic field` may occur more frequently than the phrase `magnetic resonance` in those articles from science collection. In a different collection of articles - for example a collection on nuclear medicine - the phrase `magnetic resonance` might be much more frequent than `magnetic field`. In the articles on nuclear medicine we can expect that if the word `magnetic` is present, then it is very likely that it will be followed by the word `resonance` but not the opposite. In this example, the directionality in word co-occurrences would suggest a specific relationship between the meanings of the two words in the context of nuclear medicine. In this example the `magnetic resonance` expression has the title keyword `magnetic` which has a broader meaning, which assimilates the meaning of the narrower concept of `resonance`. This relationship can be modeled as the broader concept "magnetic" includes the narrower concept "resonance".

In the thesaurus, title-descriptor associations are modeled by fuzzy binary relationships. A relationship in which two descriptors are somewhat related is called "similarity". A relationship in which the meaning of the descriptor keyword is included in the meaning of the title keyword, as in a latter example, is called "inclusion". In the following the formal definitions of fuzzy binary relationships will be presented.

3.2.2.1 General Definition

Given collections of the objects in X and Y the fuzzy binary relation R is defined on all possible pairs $(x, y) \in X \times Y$ as:

$$R = \{ \quad [\,(x, y), m_R(x, y)\,] \quad | \quad (x, y) \in X \times Y \quad \}. \tag{3.3}$$

The fuzzy binary relation (FBR) by definition 3.3 is itself a fuzzy set defined with pairs of objects. The membership function $m_R(x, y)$ for a particular pair x and y can be interpreted as the strength of the relationship between those objects. The nature of this relationship is encoded by the definition of $m_R(x, y)$, which can take a variety of forms [Zimmermann, 1996].

3.2.2.2 Definition of Similarity and Inclusion FBR

In building thesaurus the relationships between descriptors are defined as fuzzy binary relationships [Miyamoto, 1990]. Let D represent a set of documents $d_j, \quad j \in [1, \ldots, N_D]$. Let W be a set of the descriptors $w_i, \quad i \in [1, \ldots, N_W]$ of the documents in D. Let $u(w)$ be the operator returning all documents indexed by the descriptor w. Similarly $u(v)$ is the operator returning all documents indexed by the descriptor v. The FBR R defined between v and w is:

$$R = \{ \quad [\,(v, w), m_R(v, w)\,] \quad | \quad (v, w) \in W \times W \quad \}. \tag{3.4}$$

The membership function defining the **similarity** relationship between descriptors w and v is:

$$m_{R_{similarity}}(v, w) = \frac{|u(v) \cap u(w)|}{|u(v) \cup u(w)|}. \tag{3.5}$$

In the membership function given by Equation 3.5 $|.|$ defines the cardinality of the set, $\cap$ defines the intersection of the sets and the $\cup$ defines the union of the sets. The numerator $|u(v) \cap u(w)|$ in $m_{R_{similarity}}(v, w)$ represents the number of documents indexed by both descriptors v and w. The membership function defining the **inclusion** relationship between descriptors w and v is:

$$m_{R_{inclusion}}(w, v) = \frac{|u(v) \cap u(w)|}{|u(v)|}. \tag{3.6}$$

The membership function $R_{inclusion}$ encodes the relationship between w and v in which the term w is a broader concept than v. Figure 3.6 illustrates the abstract relationships between the concepts defined by the descriptors w and v. The areas associated with $u(w)$ and $u(v)$ indicate

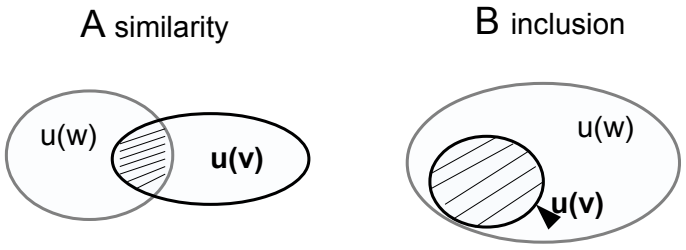


FIGURE 3.6: Graphical Representation of the Relationship Between the Descriptors. Panel A Depicts the Similarity Relationships in which the Descriptors w and v are Somewhat Related. Panel B Shows Inclusion Relationships in which Descriptor v is Fully Included in Title w.

the number of documents associated with each descriptor. The region of the intersection indicates the number of documents indexed by both: the title keyword w and the descriptor v. The left panel A in Figure 3.6 shows the schematics of the $R_{similarity}$ relationship. The strength of the relation between w and v is proportional to the area of the intersection of $u(w)$ with $u(v)$ (representing common documents indexed by both). This relationship is defined by Formula 3.5. The value of this relationship is computed as a fraction of the area of the intersection of $u(w)$ and $u(v)$ versus the union of $u(w)$ and $u(v)$. The $m_{R_{similarity}}(w, v) = 1$ occurs only if $u(w) = u(v)$. It means that all documents in the collection are indexed by both v and w.

The right panel B in Figure 3.6 illustrates the $R_{inclusion}$ relationship defined by Formula 3.6 for the pair of descriptors (w, v). Its value is computed as a fraction of the area of the intersection of $u(w)$ and $u(v)$ versus the area of $u(v)$. The $m_{R_{inclusion}}(w, v) = 1$ occurs only if $u(w) \supset u(v)$ meaning that the document set indexed by w fully contains **ALL** documents indexed by v. This also implies that the descriptor w in articles is always observed with the descriptor v, but not the other way around. The inclusion relationship is directional and models a situation such

that a broader concept includes narrower concepts. In the former example the broader concept of the title keyword `magnetic` included the narrower concept `resonance`. It can be expressed as $resonance \subset magnetic$. Added features of directionality in relationships between concepts facilitate formal inferences of links between the concepts by fuzzy operations.

3.2.3 Modeling the Relationships between Phenotype and Gene Functions by FBR

In this section principles of fuzzy set theory will be applied to provide a mathematical foundation for the literature-based algorithm. MeSH terms and GO annotations will be characterized as fuzzy sets. Fuzzy binary relationships will be defined on those sets. Fuzzy binary operations applied to these relationships will be explained. These operations provide a theoretical foundation for the actual computational procedure which is utilized to relate phenotype definitions to gene functions described by GO annotations.

Generally, phenotypes represent a broad scope of processes. This scope is broader than specific activities carried out by molecular entities such as drugs and chemicals. For example, a disease presents a broader concept than a drug. Or, a biological process is a broader concept than proteins taking part in that process. Similarly, it is assumed that gene functions are more likely to be affected by chemicals and drugs than vice versa. This directionality of the relationships is modeled by fuzzy binary inclusion relationships in the developed prioritization algorithm. The inferences in the algorithm are based on the following semantic relationships: (i) the phenotype, defined by *MeSH G term* (or of either category), is a broader concept than the chemical, defined by *MeSH D term* and (ii) the chemical is a broader concept than the gene function, defined by the *GO annotation*. The inclusion FBR that relate the *MeSH D term* to the *MeSH G term* and the *GO annotation* to the *MeSH D term* are defined as follows.

3.2.3.1 Definitions of FBR between MeSH terms and GO annotations

Let us denote MeSH G annotations as g_j, $j \in (1 \dots n_g)$ in which j refers to a particular *MeSH G term*. Similarly, let us denote MeSH D annotations by g_k, $k \in (1 \dots n_d)$. A subset of N_G MEDLINE articles annotated by the specific MeSH G annotation j is denoted by G_j. Similarly, a subset of N_D MEDLINE articles annotated by the particular MeSH D annotation k is denoted by D_k. We define the following fuzzy binary relation R_{GD} between the two MeSH terms (g_j, d_k) as:

$$\{[\,(g_j, d_k), m_{R_{gd}}(g_j, d_k)\,] \quad | \quad (g_j, d_k) \in G_j \times D_k\}, \tag{3.7}$$

with the membership function

$$m_{R_{gd}}(g_j, d_k) = \frac{|G_j \cap D_k|}{|D_k|}. \tag{3.8}$$

The brackets $|\cdot|$ in Equation 3.8 denote the cardinality (the number of elements in a set) of the intersection $|G_j \cap D_k|$ of the two sets. The intersection represents a set of the articles, annotated by both terms (g_j, d_k). The membership function in Equation 3.8 models the degree of inclusion $d_k \subseteq g_j$ of the concept D (chemical) into the concept G (phenotype).

In a similar way the inclusion relationship between the GO annotation and the concept D is defined. The difference in this case is that the relationships between the concepts are established through the genes instead of articles as in the definitions by Equations 3.7 and 3.8. To compute these relationships the databases *gene2go* and *gene2pubmed* of NCBI [Maglott et al., 2011] are used. The *gene2go* database stores GO annotations of genes. The *gene2pubmed* database stores associations between genes and PubMed articles. The GO annotations of a particular gene can be related to the MeSH D terms of the articles associated with that gene by using data in these databases.

Let us denote GO annotations by o_i, and i refers to a particular annotation. Let us denote by GO_i a subset of total N_{GO} genes annotated by the specific GO annotation o_i. Let us denote

by GD_k a subset of N_{GD} genes that were associated with articles, annotated by the MeSH D term d_k. The fuzzy binary relation R_{DO} between the *MeSH D term* and the *GO annotation* is defined as:

$$\{[\,(o_i, d_k), m_{R_{do}}(o_i, d_k)\,] \quad | \quad (o_i, d_k) \in GO_i \times GD_k\}, \tag{3.9}$$

with membership function

$$m_{R_{do}}(o_i, d_k) = \frac{|GO_i \cap GD_k|}{|GO_i|}. \tag{3.10}$$

The degree of inclusion of the *GO annotation* into the *MeSH D term* is given by the membership function in Equation 3.10.

Fuzzy sets, defined by Equations 3.7 and 3.9, can be interpreted as statements: *GO annotation* is included in *MeSH D term* and *MeSH D term* is included in *MeSH G term*. These statements can be combined to make the following inference:

if
the *GO annotation* is included in the *MeSH D term* to some degree
and
the *MeSH D term* is included in the *MeSH G term* to some degree
then
the *GO annotation* is included in the *MeSH G term* to some degree.

The degree of inclusion of the *GO annotation* in the *MeSH G term* is computed by applying maximum composition operation to the fuzzy sets defined by FBR in Equations 3.7 and 3.9.

3.2.3.2 Maximum Composition Operation

Maximum composition operation is applied on the fuzzy sets represented by fuzzy binary relationships [Zimmermann, 1996]. Maximum composition operation applied to FBRs defined by Equations 3.7 and 3.9 results in a fuzzy binary relationship whose membership function is:

$$R_{GD} \circ R_{DO} = \{[(o_i, g_j), \max(m_{R_{gd}}(d_k, g_j) * m_{R_{do}}(o_i, d_k))] \,|\, o_i \in GO_i,\ d_k \in GD_k,\ g_j \in G_j\} \tag{3.11}$$

Equation 3.11 defines a new fuzzy set which represents the relationship: the *GO annotation* is included in the *MeSH G term* to some degree. The degree of inclusion is given by the value of the membership function $\max(m_{R_{gd}}(d_k, g_j) * m_{R_{do}}(o_i, d_k))$ in Equation 3.11. This value quantifies the strength of the connection between these two concepts. Very thorough treatment of maximum composition operation can be found in reference [Zimmermann, 1996] on pages 69-77. Maximum composition operation underlies computations in the developed prioritization algorithm and will be explained in more detail.

Figure 3.7 provides a graphical representation of several fuzzy binary relationships between the phenotype, chemicals and functional gene annotations for the "**Cell Fusion**" phenotype illustrated earlier in Figure 3.2. These relationships will be used to demonstrate the computations performed by maximum composition operation. In this diagram the relationships between the

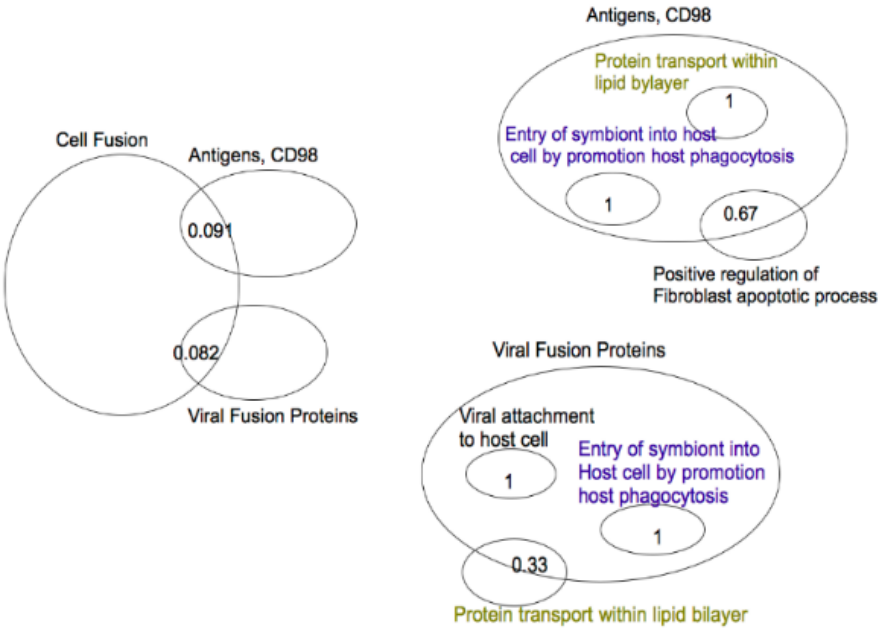


FIGURE 3.7: Graphical Representation of Inclusion Fuzzy Binary Relationships and their Values between the Terms for the **Cell Fusion** Phenotype.

concepts are represented by areas of the intersection of the ellipses that in an abstract way represent the size of the document domains indexed by these concepts. The numbers in these

intersection areas represent the FBR membership values computed by Equations 3.8 and 3.10 indicating a degree of inclusion in each individual case.

In this diagram the broader "**Cell Fusion**" phenotype is related to "**Antigens, CD98**" and "**Viral Fusion Proteins**". The chemical "**Antigens, CD98**" is related to GO annotations "**entry of symbiont into host cell by promotion host phagocytosis**", "**protein transport within lipid bilayer**" and "**positive regulation of fibroblast apoptotic process**". The concept "**Viral Fusion Proteins**" has similar relations.

The chemical"**Viral Fusion Proteins**" fully includes the GO annotation "**viral attachment to host cell**" having the FBR value of this relationship equal to 1. This means that all genes, associated with the articles annotated by "**Viral Fusion Proteins**", were annotated by the "**viral attachment to host cell**" functional GO annotation. The computation of the maximum composition operation is illustrated by an analysis of the expression $\max(m_{R_{gd}}(d_k, g_j) * m_{R_{do}}(o_i, d_k))$ in Equation 3.11 using terms and values shown in Figure 3.7.

The expression $m_{R_{gd}}(d_k, g_j)$ is the membership function of the fuzzy set defined by Equation 3.7 on the pairs (*MeSH D Term, MeSH G Term*). It represents the fuzzy binary relationship between the phenotype and chemicals. The expression $m_{R_{do}}(o_i, d_k)$ is the membership function of the fuzzy set defined by Equation 3.9 on the pairs (*GO annotation, MeSH D Term*). It represents the fuzzy binary relationship between the chemicals and functional GO annotations. Let us assume that $j = 1$, $k = 1, 2$ and $i = 1, 2$. Maximum composition operation is applied to compute the value of the fuzzy binary relationships between the phenotype "**Cell Fusion**" and GO annotations "**entry of symbiont into host cell by promotion host phagocytosis**" and "**protein transport within lipid bilayer**". Table 3.4 summarizes the notation, expressions and steps that are involved in the computations. Information in Table 3.4 complements the illustration of the maximum composition operation schematically presented in Figure 3.8.

The membership values expressing the strength of relationships between the "**Cell Fusion**"

TABLE 3.4: Steps to Compute Maximum Composition Operation on Relationships in Figure 3.7

Expression	**Explanation**
g_1	Cell Fusion
d_1	Antigens,CD98
d_2	Viral Fusion Proteins
o_1	protein transport within lipid bilayer
o_2	entry of symbiont into host cell by promotion host phagocytosis
$m_{R_{gd}}(g_1, d_1) = 0.091$	m(Cell Fusion, Antigens,CD98)
$m_{R_{gd}}(g_1, d_2) = 0.082$	m(Cell Fusion, Viral Fusion Proteins)
$m_{R_{do}}(o_1, d_1) = 1$	m(protein transport within lipid bilayer, Antigens,CD98)
$m_{R_{do}}(o_2, d_1) = 1$	m(entry of symbiont into host cell by promotion host phagocytosis, Antigens,CD98)
$m_{R_{do}}(o_1, d_2) = 0.33$	m(protein transport within lipid bilayer, Viral Fusion Proteins)
$m_{R_{do}}(o_2, d_2) = 1$	m(entry of symbiont into host cell by promotion host phagocytosis, Viral Fusion Proteins)
$\max(m_{R_{gd}}(g_1, d_.) * m_{R_{do}}(d_., o_1)) = \max(0.091 \times 1, 0.082 \times 0.33) = 0.091$	m(Cell Fusion, protein transport within lipid bilayer)
$\max(m_{R_{gd}}(g_1, d_.) * m_{R_{do}}(d_., o_2)) = \max(0.091 \times 1, 0.082 \times 1) = 0.091$	m(Cell Fusion, entry of symbiont into host cell by promotion host phagocytosis)

and the *MeSH D terms*, are combined with the membership values, expressing the strength of the relationships between the *MeSH D terms* and the *GO annotations* by multiplying them and taking the maximum. The resulting value for the pair "**Cell Fusion**" and "**protein transport within lipid bilayer**" equals 0.091 as well as the value for the pair "**Cell Fusion**" and "**entry of symbiont into host cell by promotion host phagocytosis**". Multiplication and maximum operations on fuzzy binary relationships result in the strongest relationships on pairs. Therefore, the computation will deduce the strongest relationships between a phenotype defined by *MeSH G term* and the *GO annotations*. The values of these relationships will be assigned as weights to GO annotations.

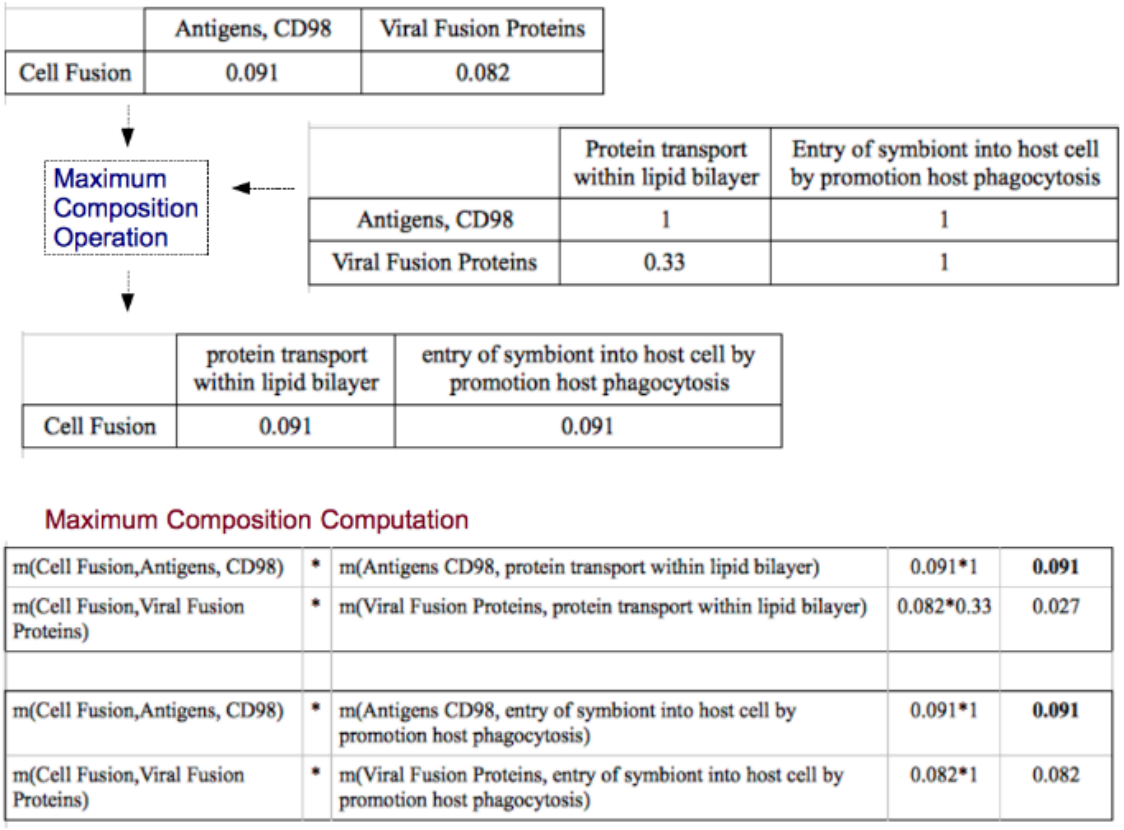


FIGURE 3.8: Maximum Composition Computation for **Cell Fusion** Example in Figure 3.7.

3.2.3.3 Inference of Relationships

In previous sections the mathematical foundation of the literature-based algorithm was presented. Here the overall process of the inference is summarized graphically. The scheme representing the inference of the links between the phenotype and the GO annotations is illustrated in Figure 3.9. This scheme of the inference in literature-based algorithm will assist us in further discussion of advantages and limitations of this algorithm in Section 3.5.

The rectangles enclose the values of membership functions that express the strength of the connections between terms. As shown in Figure 3.9, each GO annotation is assigned a numerical value expressing the strength of a connection between the phenotype and GO annotation. This value is represented in bold and is computed by the maximum composition operation on individual fuzzy sets as explained in Section 3.2.3.2. Other values in the square brackets accompanying

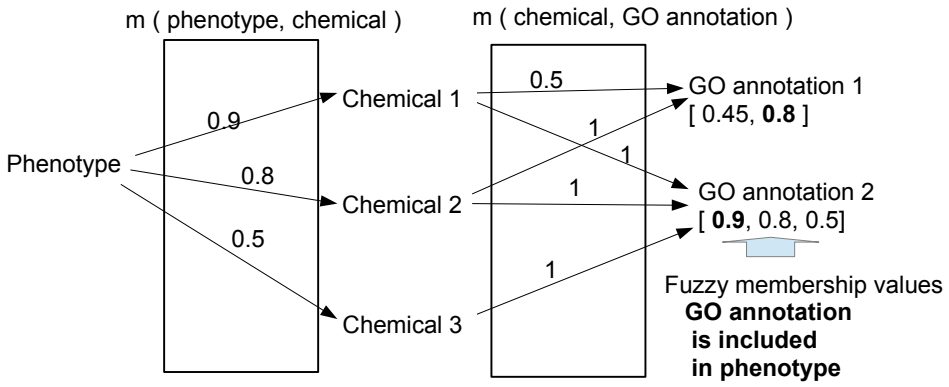


FIGURE 3.9: Schema of the Inference of Fuzzy Membership Values

the GO annotations represent the strengths of competing connections arising from linking the unique terms in the pair (phenotype GO annotation) through different chemicals (Equation 3.10). The maximum value in bold represents the final value assigned to the GO annotation by the algorithm which is used to rank genes with respect to the phenotype. The G2D gene prioritization system uses this approach as well, but only infers links between human diseases and genes in specific regions of the genome. Somewhat similar ideas of making links between the concepts are employed in the general purpose tool Arrowsmith [Smalheiser and Swanson, 1998].

3.2.3.4 Thresholds in α-cuts

Fuzzy sets with a bounded membership function are called α-cuts [Perez-Iratxeta et al., 2002b, Zimmermann, 1996]. Bounded membership means that some pairs of objects from the fuzzy set on which the membership function is defined do not belong to that fuzzy set because their membership value is below some threshold α. α-cuts represent fuzzy sets with stringent membership by bounding the membership functions in the Equations 3.10 and 3.8 from below:

$$m_{R_{gd}}(g_j, d_k) \geq t_{GD}, \quad m_{R_{do}}(o_i, d_k) \geq t_{DGO}. \tag{3.12}$$

The specific pairs of concepts (o_i, d_k) and (g_j, d_k) whose membership values are below the established thresholds are not included in the α-cuts. These pairs are not used in the maximum composition operation given by Equation 3.11 to compute the link between the GO annotation and the phenotype defined by the MeSH G term. It means that some intermediate connections through the chemicals defined by the MeSH D terms are not included in the inference. Some intermediate connections which have very low fuzzy membership values in their respective fuzzy sets may represent weak links that happened by chance. By varying the thresholds in α-cuts we can control how much information we use in linking the GO annotation to the phenotype.

The thresholds t_{DGO} and t_{GD} are used to filter spurious and noisy links. t_{GD} is called the D threshold and t_{DGO} is called the GO threshold. These parameters establish minimum values of connection strength between *phenotype* and *MeSH D* and between *MeSH D* and *GO annotation* in order for them to be included in linking the phenotype to the GO annotation. Table 3.5 represents a chain of links between the concepts and the parameters.

TABLE 3.5: Thresholds on Relationships between Concepts in Connecting Phenotype and GO Annotation

Phenotype	$\supseteq$	**MeSH D Term**	$\supseteq$	**GO Term**
MeSH term in A, B, C, D, G categories	t_{GD} threshold	Drugs and Chemicals	t_{DGO} threshold	gene annotation

3.3 Validation of Gene Prioritization Algorithms

In an ideal assessment of a prioritization tool, many runs would be carried out followed by experimental identification of the false positives rate. However, such validation is not practically possible. Therefore, a benchmark was created to test whether the developed prioritization methods were able to produce meaningful associations between genes and phenotypes of interest.

3.3.1 The benchmark

The purpose of the benchmark was to evaluate how genes that are known to be involved in particular phenotypes are prioritized by the created algorithms. The test set in the benchmark consisted of 110 phenotypes with associated genes from the Mouse Genome Informatics (MGI) database [Eppig et al., 2012]. The genes in these sets have been experimentally linked to their respective phenotypes by knock out experimental evidence. MGI phenotypes have been mapped to MeSH terms by the Mammalian Phenotype ontology (MP) [Smith and Eppig, 2012]. Among the studied phenotypes 95 corresponded to the MeSH terms of category G (Biological Processes and Phenomena), and others to category C (Diseases) which total 15.

3.3.2 Criteria to estimate performance of the prioritization algorithms

Benchmarking studies for disease genes suggest that if a small candidate set appears in the Top 5-15% of the ranked whole genome, then the prioritization captured useful information [Bornigen et al., 2012, Moreau and Tranchevent, 2012]. For the human genome the 5-15% consists of about 1,250-3,750 protein coding genes. Referring to the sizes of the gene lists used in that study [Bornigen et al., 2012], a small set of candidate genes contains about $10 - 18$ genes. Therefore, a well-performing gene prioritization tool is anticipated to place 10 to 18 candidate genes in the Top 5-15% of the ranked whole genome.

In the created benchmark the developed algorithms ranked the $24,732$ annotated genes of the mouse genome. The criteria used in this benchmark consist of estimates of how many genes from the benchmark sets appear within the Top 1% or 5% of the ranked genome. For the mouse genome the 1-5% consists of about 247-1,236 genes. Sizes of the gene sets in the created benchmark vary from 10 to 512 genes. Recall in the Top 1% of the ranked genome is a very stringent requirement compared to the criteria of 5-15% used by other studies [Moreau and Tranchevent, 2012]. The performance of the literature-based part of the prioritization algorithm was assessed

in the benchmark by permutation testing. The performance of the algorithm component which is based on the training genes was assessed in two- and five-fold cross-validation.

3.3.3 Method of Prioritization Performance Assessment Used in the Literature-Based Part of the Algorithm

The test cases of the benchmark consist of a phenotype, characterized by a MeSH keyword, and a set of genes that are already known to be related to that phenotype through experimental evidence. For each test case, a ranking of the whole genome was obtained with respect to the phenotype. Then, the number of genes in that test case ranked within the Top 1% and 5% of the ranked whole genome (true positives) was computed. A recall (a fraction of the true positives in the gene set of that test case) was estimated within the Top 1% and 5% of the ranked whole genome for all test cases of the benchmark.

In order to test whether an observed recall could be as high by chance, an empirical permutation test was applied. For each test case $10,000$ random rankings of the whole genome were generated. The number of random rankings was counted in which an equivalent or higher recall within the Top 1% and 5% of the ranked genome list was observed. This fraction of the equivalent recall in $10,000$ random rankings can be interpreted as an empirical p value. Empirical p values were estimated for each test case of the benchmark for both thresholds: the Top 1% and 5% of the ranked genome. The recalled genes in the test case were deemed significant predictions if the empirical p value for that set was less than or equal to 0.05.

The literature-based part of the prioritization algorithm depends on two parameters: the D threshold and the GO threshold. These parameters determine the minimum values of the fuzzy membership in fuzzy sets defined on MeSH terms and GO annotations. The results of the benchmark will be discussed with respect to the fixed D and GO thresholds 0.001 and 0.01 respectively. With these values there were significant predictions in 87.3% of the benchmark cases

within the Top 1% of the ranked genome. Selection of the parameters is explained in Section A.3 in Supplementary Material.

3.3.4 Method of Cross-Validation Prioritization Performance Assessment Used in the Algorithm Part Based on Training Genes

The gene sets of the benchmark were used to test the prioritization performance of the algorithm part based on training genes to define a phenotype. This performance was assessed by cross validation (CV). Cross-validation is a procedure in which a proportion of the data (the training set) is used to train a system. The remaining part of the data (the test set) is put aside to evaluate how well the trained system is performing on unseen data. The data split is usually repeated several times. Each time the system is trained using the training set and tested on the test set [Bornigen et al., 2012, Yu et al., 2008a]. Obtained values of a selected metrics of the system performance are averaged across all splits. The averaged value is used to assess the performance of the system.

Cross-validation was used to estimate the prioritization performance of the algorithm component that uses training genes to define the phenotype. The genes sets of the benchmark test cases were used in training and in testing. The training consisted of estimating the semantic similarity (explained in Section 3.1.3) between the pooled GO annotations of the training genes and the GO annotations of the entire ontology. A result of training is that the GO annotations of the entire ontology are assigned weights that measure their functional similarity to the training genes. Then, the whole genome is ranked according to the magnitude of the weights assigned to the gene annotations. The testing consists of estimating how many of the test genes, not used in training, are ranked within the Top 1% and 5% of the ranked genome. A metric of the performance is the recall of the test genes within the Top 1% and 5% of the ranked genome. High recall within the Top 1% would indicate a good performance of the prioritization algorithm.

Two- and five-fold cross-validation setups were used. In the two-fold CV, the gene set of the benchmark test case is split into two parts. Training takes place on one part and testing on the other. Then the training and the test sets are interchanged and the process is repeated again. In the five-fold CV, the gene set is split into five equal parts. In a real life gene prioritization scenario, a user might have only a few known genes to define a phenotype. To follow this realistic scenario one fifth of the gene set was used in training, and the remaining genes were used to estimate the recall. This was repeated for each such subdivision, five times in total.

3.3.5 Benchmark Performance in Prioritization Algorithms

This section reviews the main results obtained in the assessment of prioritization performance in the created benchmark for both parts of the prioritization algorithm.

Prioritization Performance in the Literature-Based Part of the Algorithm. The benchmark gene sets differed in size, recall and the significance of the recall. Table 3.6 presents overall significance and recall obtained in the benchmark for the phenotypes of the MeSH category G. The recall is reported as a percentage of genes from the benchmark set occurring within the

TABLE 3.6: The Recall and Significance Achieved by the Literature-Based Prioritization Algorithm in the Benchmark for the Phenotypes Defined by MeSH G Keywords.

Recall (the intervals of the recall(%) in the benchmark sets)	**The number of the Top 1% benchmark sets**	**The number of the Top 5% benchmark sets**
[0..10]	21(9)	4,(0)
(10..20]	24(24)	3,(1)
(20..30]	27(27)	10,(10)
(30..40]	10(10)	14,(14)
(40..50]	8(8)	19,(19)
(50..100]	5(5)	45,(45)
Percent of significant recall	*87.4%*	*93.7%*

Top 1% and 5% of the genome. For each interval of the recall (%) the number of benchmark sets within the Top 1% and 5% of the ranked genome is shown. Numbers in parenthesis show how many sets had significant predictions (having p value ≤ 0.05). Significant recall within the

Top 1% of the ranked genome was observed in 87.4% of the 95 benchmark gene sets. Within the Top 5% of the ranked genome there were 93.7% benchmark sets with significant predictions. Figure 3.10 represents how the recall differed across gene sets of various sizes in the benchmark cases which had significant predictions. The x axis in the Box-Whisker plots in Figure 3.10

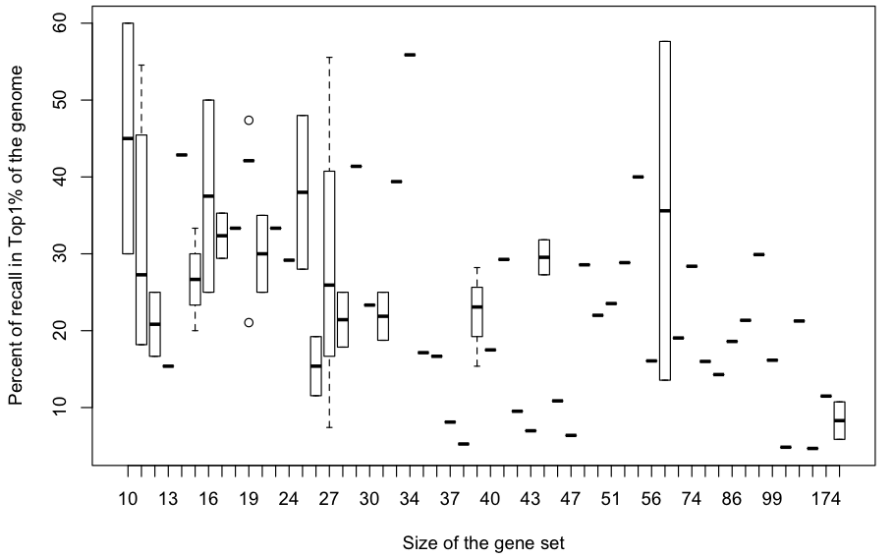


FIGURE 3.10: Recall within the Top 1% of the Whole Ranked Genome versus the Gene Set Size in the Benchmark Test Cases Having Significant Predictions. On the y Axis the Percentage of Recalled Genes within the Top 1% of the Genome is Shown.

represents the size of the gene set. The y axis represents the percentage of recalled genes within the Top 1% of the genome for the gene set of a fixed size. There was no relationship between the size of the gene set and the recall. The median size of the gene set in the benchmark was 32. The median recall in the Top 1% of the ranked genome was 23.53%. In the Top 5% of the ranked genome the median recall was 51.69%. Based on those estimates on average there were $7 \sim 8$ significant predictions within the Top 1% of the ranked genome and about $16 \sim 17$ within the Top 5%. This result compares very favorably with the current standards in prioritization performance reported in literature [Moreau and Tranchevent, 2012]. The overall performance of the literature-based prioritization algorithm in the benchmark test cases is presented in Tables

3.8, 3.9 and 3.10 for large (512 - 56), moderately (52 - 32) and smaller (30-17) sized gene sets respectively. These tables summarize the algorithm performance within the Top 1% of the ranked genome. The number of recalled genes, the percentage of recall and the empirical p value obtained by permutation testing are presented for each tested phenotype represented by a MeSH G keyword and corresponding MGI phenotype. Most of the benchmark test cases had significant predictions. Very good recall of more than 50% in the Top 1% of the genome was observed in phenotypes "**Blood Coagulation**", "**DNA Repair**", "**Leukocyte Rolling**", "**Vasoconstriction**" and "**DNA Replication**". Some phenotypes without significant predictions in the Top 1% of the ranked genome are presented in Table 3.7. Several of them had improved their recall within the Top 5% of the ranked genome except the "**Autophagy**" and "**Immune System Phenomena**" phenotypes. The results for the "**Autophagy**" phenotype will be reviewed later in this Chapter.

TABLE 3.7: Phenotypes without Significant Predictions

MeSH G Term	**MGI Phenotype**	**Gene Set Size**	**Recall within the Top 1%**			**Recall within the Top 5%**		
			%	#	p	%	#	p
Immune System Phenomena	abnormal immune system physiology	107	0	0	1	1	0.93	0.99
Homeostasis	abnormal homeostasis	33	1	3.03	0.23	10	30.3	0
Respiration	abnormal respiration	22	1	4.55	0.23	5	22.73	0
Autophagy	abnormal autophagy	18	0	0	1	3	16.67	0.06
Wound Healing	abnormal wound healing	13	1	7.69	0.13	5	38.46	0

Overall, the results achieved in the benchmark are very encouraging. The benchmark relies on experimental evidence of associations between genes and phenotypes. There might be genes truly related to the phenotype; however, because of a lack of experimental evidence they are not in the test set. For this reason it is possible that the performance of the literature-based prioritization algorithm is slightly underestimated in the current benchmark.

TABLE 3.8: Benchmark Performance of the Prioritization Algorithm within the Top 1% of the Ranked Genome using MGI Gene Sets Sized 512 - 56

MeSH G Term	**MGI Phenotype**	**Gene Set Size**	**Recall#**	**Recall %**	p **value**
Organogenesis	complete embryonic lethality during organogenesis	512	55	10.74	$\leq 10^{-7}$
Genes, Lethal	complete embryonic lethality during organogenesis	512	30	5.86	$\leq 10^{-7}$
Spermatogenesis	abnormal spermatogenesis	174	20	11.49	$\leq 10^{-7}$
Fetal Development	complete lethality throughout fetal growth and development	171	8	4.68	0.0001
Bone Density	decreased bone mineral density	127	27	21.26	$\leq 10^{-7}$
Evoked Potentials, Auditory	decreased brainstem auditory evoked potential	124	6	4.84	0.0015
Immune System Phenomena	abnormal immune system physiology	107	0	0	1.0000
Hematopoiesis, Extramedullary	extramedullary hematopoiesis	99	16	16.16	$\leq 10^{-7}$
Erythropoiesis	abnormal erythropoiesis	97	29	29.9	$\leq 10^{-7}$
Hematopoiesis	abnormal hematopoiesis	89	19	21.35	$\leq 10^{-7}$
Cell Cycle	abnormal cell cycle	86	16	18.6	$\leq 10^{-7}$
Birth Weight	decreased birth weight	84	12	14.29	$\leq 10^{-7}$
Spermatogenesis	arrest of spermatogenesis	75	12	16	$\leq 10^{-7}$
Synaptic Transmission	abnormal CNS synaptic transmission	74	21	28.38	$\leq 10^{-7}$
Cell Movement	abnormal cell migration	63	12	19.05	$\leq 10^{-7}$
Whole Blood Coagulation Time	abnormal blood coagulation	59	8	13.56	$\leq 10^{-7}$
Blood Coagulation	abnormal blood coagulation	59	34	57.63	$\leq 10^{-7}$
Immune System Phenomena	abnormal immune system organ morphology	59	0	0	1.0000
Oxygen Consumption	increased oxygen consumption	56	9	16.07	$\leq 10^{-7}$
Inhibitory Postsynaptic Potentials	abnormal inhibitory postsynaptic currents	55	22	40	$\leq 10^{-7}$

TABLE 3.9: Benchmark Performance of the Prioritization Algorithm within the Top 1% of the Ranked Genome using MGI Gene Sets Sized 52 - 32

MeSH G Term	**MGI Phenotype**	**Gene Set Size**	**Recall#**	**Recall %**	p **value**
Oxidative Stress	oxidative stress	52	15	28.85	$\leq 10^{-7}$
Wound Healing	abnormal vascular wound healing	51	12	23.53	$\leq 10^{-7}$
Corpus Luteum Maintenance	absent corpus luteum	50	11	22	$\leq 10^{-7}$
Gastrulation	abnormal gastrulation	49	14	28.57	$\leq 10^{-7}$
Fetal Weight	decreased fetal weight	47	3	6.38	0.0200
Necrosis	necrosis	46	5	10.87	$\leq 10^{-7}$
Mitosis	abnormal mitosis	44	14	31.82	$\leq 10^{-7}$
Cell Adhesion	abnormal cell adhesion	44	12	27.27	$\leq 10^{-7}$
Immune System Phenomena	abnormal immune system morphology	43	0	0	1.0000
Brain Waves	abnormal brain wave pattern	43	3	6.98	0.0087
Gametogenesis	abnormal gametogenesis	42	4	9.52	0.0011
Lymphopoiesis	abnormal lymphopoiesis	41	12	29.27	$\leq 10^{-7}$
Cell Death	abnormal cell death	40	7	17.5	$\leq 10^{-7}$
Blood Circulation	abnormal blood circulation	40	7	17.5	$\leq 10^{-7}$
Skin Pigmentation	abnormal skin pigmentation	39	9	23.08	$\leq 10^{-7}$
Myelopoiesis	abnormal myelopoiesis	39	11	28.21	$\leq 10^{-7}$
Cell Cycle Checkpoints	abnormal cell cycle checkpoint function	39	6	15.38	$\leq 10^{-7}$
Evoked Potentials, Auditory	absent brainstem auditory evoked potential	38	2	5.26	0.0500
Cell Nucleus Shape	abnormal cell nucleus morphology	37	3	8.11	$\leq 10^{-7}$
Immune System Phenomena	abnormal immune system cell morphology	36	0	0	1.0000
Gluconeogenesis	abnormal gluconeogenesis	36	6	16.67	$\leq 10^{-7}$
Cell Differentiation	abnormal cell differentiation	35	6	17.14	$\leq 10^{-7}$
DNA Repair	DNA repair	34	19	55.88	$\leq 10^{-7}$
Homeostasis	abnormal homeostasis	33	1	3.03	0.2300
Chromosomal Instability	chromosomal instability	33	13	39.39	$\leq 10^{-7}$
Cell Proliferation	abnormal cell proliferation	32	6	18.75	$\leq 10^{-7}$
Aneuploidy	aneuploidy	32	8	25	$\leq 10^{-7}$

TABLE 3.10: Benchmark Performance of the Prioritization Algorithm within the Top 1% of the Ranked Genome using MGI Gene Sets Sized 30 - 17

MeSH G Term	**MGI Phenotype**	**Gene Set Size**	**Recall#**	**Recall %**	p **value**
Pregnancy	abnormal mammary gland growth during pregnancy	30	7	23.33	$\leq 10^{-7}$
Long-Term Potentiation	abnormal long term potentiation	30	7	23.33	$\leq 10^{-7}$
Hematopoiesis	abnormal embryonic hematopoiesis	30	7	23.33	$\leq 10^{-7}$
Pain Threshold	abnormal pain threshold	29	12	41.38	$\leq 10^{-7}$
Exercise	abnormal exercise endurance	28	5	17.86	$\leq 10^{-7}$
Estrous Cycle	abnormal estrous cycle	28	7	25	$\leq 10^{-7}$
Myoclonus	myoclonus	27	2	7.41	0.0300
Leukocyte Rolling	abnormal leukocyte tethering or rolling	27	15	55.56	$\leq 10^{-7}$
Bone Remodeling	abnormal bone remodeling	27	7	25.93	$\leq 10^{-7}$
Sperm Motility	abnormal sperm motility	26	3	11.54	0.0027
Chromosome Breakage	chromosome breakage	26	5	19.23	$\leq 10^{-7}$
Pharmacokinetics	abnormal xenobiotic pharmacokinetics	25	12	48	$\leq 10^{-7}$
Estrous Cycle	absent estrous cycle	25	7	28	$\leq 10^{-7}$
Erythropoiesis	abnormal embryonic erythropoiesis	24	7	29.17	$\leq 10^{-7}$
Respiration	abnormal respiration	22	1	4.55	0.2300
Synaptic Transmission	abnormal synaptic transmission	21	7	33.33	$\leq 10^{-7}$
Parturition	abnormal parturition	20	5	25	$\leq 10^{-7}$
DNA Methylation	DNA methylation	20	7	35	$\leq 10^{-7}$
Vasodilation	abnormal vasodilation	19	9	47.37	$\leq 10^{-7}$
Platelet Activation	abnormal platelet activation	19	8	42.11	$\leq 10^{-7}$
Immunity, Innate	abnormal innate immunity	19	8	42.11	$\leq 10^{-7}$
Drinking	decreased drinking behavior	19	4	21.05	$\leq 10^{-7}$
Bone Resorption	decreased bone resorption	19	8	42.11	$\leq 10^{-7}$
Muscle Development	abnormal muscle development	18	6	33.33	$\leq 10^{-7}$
Autophagy	abnormal autophagy	18	0	0	1.0000
Anovulation	anovulation	18	6	33.33	$\leq 10^{-7}$
Neovascularization, Physiologic	abnormal physiological neovascularization	17	5	29.41	$\leq 10^{-7}$
Chemotaxis	abnormal macrophage chemotaxis	17	6	35.29	$\leq 10^{-7}$

Prioritization Performance of the Algorithm Part based on Training Genes. The performance of the prioritization algorithm part based on phenotype definition by training genes was tested on 87 benchmark gene sets. It was required that the training set in cross-validation contains at least 6 genes. In each cross-validation step recall within the Top 1% and within the Top 5% of the ranked genome was computed. These recall values were averaged. Summary of the results obtained in five-fold and two-fold CV is presented in Tables 3.11 and 3.12 respectively.

Median estimates of gene set sizes and recall from Tables 3.12 and 3.11 offer insight into the

TABLE 3.11: Average Recall (%) in Five-Fold Cross Validation of Gene Prioritization Algorithm Based on Training Genes within the Top 1% and the Top 5% of the Whole Ranked Genome

	Recall within the Top 1%	**Recall within the Top 5%**	**Training Genes #**	**Test Genes #**
Minimum	11.43	22.79	6	24
1st Quartile	20.61	41.20	8	32
Median	24.69	46.72	10	41
Mean	25.55	46.86	16	63
3rd Quartile	30.16	52.31	18	70
Maximum	44.43	68.65	122	490

average performance of our system in the 87 tested benchmark cases.

If we have 10 known genes in the training set, then we can expect that about 24% of the good candidates will be identified in our candidate list within the Top 1% of the genome. Within the Top 5% of the genome about 47% of the good candidates will be identified. If we have more training genes as in the two-fold CV setting, then on average with the 26 training genes we can anticipate finding about 37% and 59% of the good candidates in the respective Top 1% and Top 5% of the ranked genome. These cross validation results are promising, but nevertheless they are specific to the tested gene sets.

Observation of the same performance cannot be guaranteed if using other gene sets because the heterogeneity of the tested gene sets in this CV study was not accounted for. The reported estimates can be regarded as only a rough approximation of the expected performance of the

prioritization algorithm, based on the training genes for the phenotype definition.

It was observed that bigger training sets lead to better recall. Median values in Tables

TABLE 3.12: Average Recall (%) in Two-Fold Cross Validation of Gene Prioritization Algorithm Based on Training Genes within the Top 1% and the Top 5% of the Whole Ranked Genome

	Recall within the Top1%	**Recall within the Top5%**	**Test and Training Genes #**
Minimum	21.91	43.87	16
1st Quartile	32.12	53.28	20
Median	37.50	59.38	26
Mean	38.60	59.60	39
3rd Quartile	43.75	65.27	44
Maximum	61.67	81.67	306

3.11 and 3.12 indicate this trend. This result is not surprising because the larger training set has more information.

The minimum attainable recall within the Top 1% of the genome varied from 10% to 20% if the algorithm was trained by a minimum number of genes from 6 to 9. The gene sets having average recall below the first Quartile ≤ 20 % within the Top 1% of the whole genome perform poorly compared to the recall in other gene sets. These gene sets are shown in Table 3.13. In the five-fold setting the size of the training set is much smaller than the training set in the two-fold CV setting. As can be seen from Table 3.13 for all phenotypes the recall in the two-fold CV always improved, implying that with larger training set sizes we may expect better recall, because more genes are available for training in the two-fold cross validation setting.

Analysis of the improvement in recall based on an increase of the training set size going from the five-fold to two-fold CV setting in Table 3.13 shows that in our benchmark a 2.5-fold increase in gene set size leads to 1.5− to 3−fold increase in recall. A greater increase in recall is observed if initially there were fewer genes for training as in the gene sets associated with the phenotypes "**cell adhesion**", "**abnormal gastrulation**" and "**increased bleeding time**" (Table 3.13).

TABLE 3.13: Phenotypes Represented by Gene Sets with Poor Recall (%) within the Top 1% and within the Top 5% of the Whole Genome

MGI Phenotype	**Five-Fold CV**				**Two-Fold CV**		
	Recall in Top 1%	**Recall in Top 5%**	**Test Set Size**	**Training Set Size**	**Recall in Top 1%**	**Recall in Top 5%**	**Gene Set Size**
impaired glucose tolerance	19.58	37.06	116	29	34.11	52.93	73
abnormal gametogenesis	19.57	33.74	40	10	45.24	71.43	25
abnormal hematopoiesis	19.1	41.58	75.2	18.8	29.93	55.69	47
abnormal cell adhesion	15.9	45.41	36	9	35.2	55.59	23
abnormal spermatogenesis	15.56	30.94	168.8	42.2	24.81	43.87	106
complete embryonic lethality during organogenesis	14.91	36.17	489.6	122.4	21.91	46.22	306
increased bleeding time	14.91	53.72	53.6	13.4	41.44	57.98	34
arrest of spermatogenesis	14.63	34.84	71.2	17.8	28.18	53.99	45
abnormal gastrulation	14.28	40.81	39.2	9.8	40.17	64.92	25
enhanced long term potentiation	12.8	34.07	39.2	9.8	28.26	55.98	25
cell adhesion	11.43	22.79	36	9	34.09	47.73	23

3.4 Application of the Developed Prioritization Algorithm

Prioritization application in this section concerns the prioritization of genes with respect to the cell fusion phenotype, which is a non-disease phenotype. This section intends to highlight the utility of the algorithm in acquiring insights on which genes might be the strongest candidates related to the phenotype of interest. Prioritizing genes with respect to cell fusion illustrates a situation in which not much information is available about the relationships between those genes and the phenotype of interest. Another application of the developed algorithm presented in this section concerns the prioritization of genes reported in literature in relation to cancer

phenotypes.

3.4.1 Cell Fusion Example

Identifying genes related to a biological process of interest which has not been well characterized is challenging. The myoblast cell fusion phenotype presents one such example. During embryonic and adult myoblast fusion in the mouse, differentiated myoblasts elongate and migrate towards other myoblasts. The myoblasts adhere to each other and fuse, forming nascent myotubes. Later, more myoblasts fuse with the nascent myotubes, giving rise to a mature myotube [Dittmar and Zanker, 2011]. Candidate genes critical for myoblast cell fusion induced by p38α MAPK signaling (total 214 genes) were identified by microarrays [Liu et al., 2012]. Among them Cd53 was identified as a novel mediator of myoblast cell fusion. The roles of most of other genes in this process were not characterized. The p38α-induced genes were analyzed by the prioritization algorithm with respect to the processes and structures related to myoblast cell fusion. These phenotypes are "**Cell Fusion**", "**Membrane Fusion**", "**Cell Adhesion**", "**Muscle Cell Development**", "**Myoblasts**", "**Actin Cytoskeleton**", "**Extracellular Matrix**" and "**Intercellular Junctions**". Table 3.14 summarizes the top scoring genes for each phenotype.

The membrane proteins Cd81 and Cd9 are known to be related to myoblast cell fusion. They are involved in cell-cell adhesion processes and regulate beta1 integrin signaling, which plays an important role in myoblast cell fusion [Dittmar and Zanker, 2011]. We defined the phenotype by the known Cd9 and Cd81 genes and ranked the p38a-induced genes by functional similarity to the phenotype defined by the known genes.

The top scoring genes with respect to the "**Membrane Fusion**" phenotype were very similar to the top genes with respect to the phenotype defined by the context genes Cd81 and Cd9. The novel mediator of myoblast cell fusion Cd53 is present in both subsets of the top scoring genes. This result is in line with the knowledge about Cd53 being a part of the tetraspanin

TABLE 3.14: The p38α Induced Genes Critical to Myoblast Cell Fusion that Rank within the Top 1% of the Whole Genome with Respect to the "**Cell Fusion**" Related Phenotypes

Known genes Cd81, Cd9	**Cell Fusion**	**Cell Adhesion**	**Membrane Fusion**	**Muscle Cell Development**	**Myoblasts**	**Actin Cyto Skeleton**	**Extra Cellular Matrix**	**Inter Cellular Junctions**
Ifitm3 Id2 Cdkn1a Myh9 Cd53	Csf1 Slc7a7 Fn1 Pla2g16 Ifi204 Edn1	Tnc Lrp1 Fn1 Csf1 Ctgf Edn1 Dapk2 Ctla2a Hbegf Vcl Ctnna1 Myh9	Ifitm3 Id2 Clic4 Cdkn1a Myh9 Cd53 Diap2	Ifi204 Tead4 Runx1 Myog Lef1 Lpar1 Ncam1 Id2 Ctgf	Ifi204 Runx1 Tead4 Myog Inhba Ctgf Lef1 Id2 Mb	Cald1 Diap2 Actn4 Myo1d Megf10 Lrp1 Capg7	Tnc Csf1 Col5a1 Col5a2 Lrp1 Ctla2a Col1a1 Edn1 Col12a1 Fn1 Phospho1 Slc4a7 Has2 Runx1	Ctnna1 Smad7 Vcl Scn5a Pkp1 Lef1 Tead4 Actn1 Tspo1

web and acting on sites of cell-cell contact [Liu et al., 2012]. Some genes may have very broad functional spectrums and are strongly related to multiple phenotypes. Such genes are Fn1, Edn1, Csf, Vcl, Lrp1, Id2, Lef1, Runx1 and Myh9. The unique genes strongly connected to the "**Cell Fusion**" phenotype are Slc7a7 (solute carrier family 7 (cationic amino acid transporter, y+ system), member 7) and Pla2g16 (phospholipase A2, group XVI). Slc7a7 is characterized as integral to the membrane with molecular function amino acid transmembrane transporter activity. Pla2g16 is characterized as integral to the membrane, cytoplasm and endoplasmic reticulum and is involved in lipid metabolic and catabolic processes. Ifi204 (interferon activated gene 204) is strongly connected to cell fusion, myoblasts and muscle development. This gene is characterized by the process of cellular response to interferon beta, which is any process that results in a change in the state or activity of a cell (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of an interferon-beta stimulus. Based on current knowledge, the developed prioritization algorithm suggested that the genes Slc7a7, Pla2g16 and

Ifi204 in the p38α-induced gene set are unique, strongly connected to cell fusion processes and might play an important role in those processes.

The 45 putative genes related to the myoblast cell fusion phenotype were identified from literature and other information sources [Liu et al., 2012]. These genes are listed in Table 3.15. They were used as training genes to complement the "**Cell Fusion**" phenotype definition in prioritizing the p38α-induced genes. Putative genes provided an independent line of evidence about the phenotype. The prioritization of candidate genes by the literature-based algorithm relies on different principles than the algorithm which uses functional similarity between genes. These different approaches complement each other.

The candidate p38α-induced genes within the Top 1% of the whole mouse genome ranked by the developed prioritization algorithm according to the best combined score are shown in Figure 3.11. The "Literature" column represents the ranking, computed by the literature-based prioritization approach. The "Context Genes" column represents the ranking computed by the algorithm using training genes for phenotype definition. In this case the training genes were the putative myoblast cell fusion genes listed in Table 3.15. The genes that were ranked high by both approaches were Myh9 (myosin heavy polypeptide 9 non-muscle), Fb1 (fibronectin 1) and Edn1 (endothelin 1). Myh9 is present within the group of training genes. Therefore, it is expected that this gene will be ranked highly by the algorithm based on training genes. Myh9, which was provided by the ranking based on the training genes, can be considered a positive control in literature-based prioritization. However, the Edn1 and Fb1 genes do not belong to the set of training genes. The top scoring GO annotation for fibronectin1 is "**calcium-independent cell-matrix adhesion**" in both approaches. The top scoring GO annotation for endothelin 1 is "**positive regulation of chemokine-mediated signaling pathway**" in literature-based prioritization and "**positive regulation of sarcomere organization**" in training gene-based prioritization. These genes can be considered good candidates with respect to the "**Cell Fusion**"

Symbol	Description	CombinedScore		Literature		Homology+		ContextGenes	
Myh9 (1)	myosin heavy polypeptide 9 non-muscle	0.0540	(32)	5.6036	(1380)	7.1547	(1762)	0.0540	(10)
Csf1 (2)	colony stimulating factor 1 (macrophage)	0.1543	(93)	0.9664	(238)	0.1543	(38)	6.1033	(1131)
Runx1 (3)	runt related transcription factor 1	0.1943	(113)	1.7095	(421)	3.1957	(787)	0.1943	(36)
Slc7a7 (4)	solute carrier family 7 (cationic amino acid transporter y+ system) member 7	0.2233	(127)	15.0039	(3695)	0.2233	(55)	67.5517	(12518)
Cacna1s (5)	calcium channel voltage-dependent L type alpha 1S subunit	0.2752	(154)	10.3951	(2560)	17.5498	(4322)	0.2752	(51)
Fn1 (6)	fibronectin 1	0.3573	(208)	0.3573	(88)	1.0070	(248)	0.4209	(78)
Pla2g16 (7)	phospholipase A2 group XVI	0.4264	(250)	12.9492	(3189)	0.4264	(105)	69.8559	(12945)
Ifi204 (8)	interferon activated gene 204	0.5847	(334)	0.5847	(144)	1.5065	(371)	6.1303	(1136)
Megf10 (9)	multiple EGF-like-domains 10	0.6530	(365)	4.8159	(1186)	8.0400	(1980)	0.6530	(121)
Scn5a (10)	sodium channel voltage-gated type V alpha	0.9174	(503)	6.9233	(1705)	11.6458	(2868)	0.9174	(170)
Edn1 (11)	endothelin 1	0.9258	(508)	7.4796	(1842)	0.9258	(228)	1.1926	(221)
Lef1 (12)	lymphoid enhancer binding factor 1	0.9444	(519)	3.7601	(926)	6.1599	(1517)	0.9444	(175)
Smad7 (13)	SMAD family member 7	1.0253	(551)	6.5132	(1604)	10.9108	(2687)	1.0253	(190)

FIGURE 3.11: Prioritization Result for p38α-induced Candidate Genes with Respect to the "**Cell Fusion**" Phenotype within the Top 1% of the Ranked Mouse Genome Ordered by Combined Score.

phenotype.

In the analyzed application the proposed algorithm facilitated the elucidation of a small set of candidate genes for further investigation with respect to a cell fusion phenotype. It also aided in the formulation of hypotheses in search of those genes.

TABLE 3.15: Putative Myoblast Cell Fusion Genes.

Symbol	Name	Evidence
ADAM12	ADAM metallopeptidase domain 12 (meltrin alpha)	myoblast fusion GO in mouse
ARF6	ADP-ribosylation factor 6	myoblast fusion GO in fly
CACNA1H	calcium channel, voltage-dependent, alpha 1H subunit	myoblast fusion GO in human
CACNA1S	calcium channel, voltage-dependent, L type, alpha 1S subunit	myoblast fusion GO in mouse
CAPN2	calpain 2, (m/II) large subunit	myoblast fusion GO in mouse
CAST	calpastatin	myoblast fusion GO in mouse
CAV3	caveolin 3	myoblast fusion GO in zebrafish
CD164	CD164 molecule, sialomucin	Richardson Traffic review
CD9	CD9 molecule	Richardson Traffic review
CDC42	WASP activation	PMID: 19443691
CDH15	cadherin 15, M-cadherin (myotubule)	Richardson Traffic review
CDON	Cdon homolog (mouse)	myoblast fusion GO in mouse
CNN3	calponin 3, acidic	myoblast fusion GO in fly Mp20 homolog
CRK	v-crk sarcoma virus CT10 oncogene homolog (avian)	myoblast fusion GO in zebrafish
DOCK1	dedicator of cytokinesis 1	myoblast fusion GO in zebrafish
DOCK5	dedicator of cytokinesis 5	myoblast fusion GO in zebrafish
DYRK1B	dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 1B	myoblast fusion GO in mouse
EHD2	Putative eps protein	PMID: 18502764
GLIS3	GLIS family zinc finger 3	myoblast fusion GO in fly lameduck homolog
IL4	interleukin 4	Richardson Traffic review
IQSEC1	IQ motif and Sec7 domain 1	myoblast fusion GO in fly schizo homolog
ITGB1	integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)	Richardson Traffic review
KCNH1	potassium voltage-gated channel, subfamily H (eag-related), member 1	myoblast fusion GO in human
KIRREL	kin of IRRE like (Drosophila)	myoblast fusion GO in fly
KIRREL3	kin of IRRE like 3 (Drosophila)	myoblast fusion GO in zebrafish
LBX1	ladybird homeobox homolog 1 (Drosophila)	myoblast fusion GO in fly
MEF2A	MADS box transcription enhancer factor 2, polypeptide A (myocyte enhancer factor 2A)	myoblast fusion GO in fly
MIB2	mindbomb homolog 2 (Drosophila)	myoblast fusion GO in fly
MYH9	myosin, heavy chain 9, non-muscle	myoblast fusion GO in mouse
MYOF	Myoferlin Richardson	Traffic review
NCKAP1	NCK-associated protein 1	myoblast fusion GO in fly kette homolog
NEO1	neogenin homolog 1 (chicken)	myoblast fusion GO in rat
NFATC2	nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 2	Richardson Traffic review
NOS1	nitric oxide synthase 1 (neuronal)	myoblast fusion GO in human
NPHS1	nephrosis 1, congenital, Finnish type (nephrin)	myoblast fusion GO in fly sticks and stones homolog
NTN3	Netrin 3	Richardson Traffic review
OLFR16	MOR23	PMID: 19922870
PTGFR	prostaglandin F receptor (FP)	Richardson Traffic review
PTGIR	prostaglandin I2 (prostacyclin) receptor (IP)	Richardson Traffic review
PTGIS	prostaglandin I2 (prostacyclin) synthase	Richardson Traffic review
RAC1	ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1)	myoblast fusion GO in zebrafish
RAC2	ras-related C3 botulinum toxin substrate 2 (rho family, small GTP binding protein Rac2)	myoblast fusion GO in fly
SIX4	sine oculis homeobox homolog 4 (Drosophila)	myoblast fusion GO in fly
TRIO	triple functional domain (PTPRF interacting)	Richardson Traffic review
VCAM1	vascular cell adhesion molecule 1	Richardson Traffic review

3.4.2 Cancer Related Genes

Cancer is one of the most devastating human diseases and has been the subject of much research. Large numbers of research articles covering cancer genomics have been published. These publications are rich data resources for the literature-based prioritization algorithm developed in this thesis. By prioritizing genes with respect to cancer domains an attempt was made to explore whether the algorithm was able to identify genes already associated with cancer subtypes. Another goal was to explore whether the information generated by the algorithm can provide novel insight into genes for which associations with cancer have not been verified.

In the case of cancer, genes accumulate a large number of mutations [Roukos, 2013]. Next generation sequencing screening may produce a vast number of genetic variants and genes. If a gene harbouring a variant was not previously reported, then our tool can be used to assess how strongly that gene is connected to the specific cancer. In a study aiming to characterize somatic alterations in colorectal cancer, a set of highly mutated genes was identified [Muzny et al., 2012]. The study indicated that the mutations identified in SOX9 were not previously related to human cancer. Highly mutated genes reported in that colorectal cancer study [Muzny et al., 2012] were prioritized with respect to the phenotype "**Colorectal Neoplasms**".

The genes of that study ranking high within the Top 1% of the whole genome are listed in Table 3.16. For each gene the strongest connection between the gene annotation and the "**Colorectal Neoplasms**" phenotype through the MeSH D term is shown in addition to the gene rank and GO annotation score. This score represents the strength of the relationship between the phenotype and the GO annotation through the MeSH D term (discussed in Section 3.2.3.2). The gene SOX9 ranks very high. For a researcher who is unsure if the SOX9 gene is the cancer gene implicated in colorectal cancer our algorithm would suggest a strong connection between SOX9 and colorectal cancer. All top ranking genes in Table 3.16 are most strongly connected to the "**Adenomatous Polyposis Coli Protein**". This chemical is a negative regulator of beta-catenin

signaling in a polyposis syndrome. The polyposis syndrome is characterized by the development of numerous adenomatous polyps in the colon and rectum of affected individuals by early adulthood. The lifetime risk of colorectal cancer in these patients reaches 100 percent by age 60.

Another cancer subtype for which the developed algorithm was applied to score genome genes

TABLE 3.16: Top Ranking Genes with Respect to the "**Colorectal Neoplasms**" Phenotype. The Strongest Connection between the Gene Annotation and the Phenotype through the MeSH D Term is Shown for Each Gene. Gene Score and Rank in the Whole Genome with Respect to"**Colorectal Neoplasms**" is Shown as Well as the Score of GO Annotation and the Connecting MeSH D Term

Gene	Description	Gene Score (Gene Rank)	GO Annotation	GO Score	MeSH D
APC	adenomatous polyposis coli	0.121 (27)	regulation of glutamate metabolic process	0.169	Adenomatous Polyposis Coli Protein
CTNNB1	catenin (cadherin-associated protein) beta 1 88kDa	0.127 (30)	mesenchyme morphogenesis	0.169	Adenomatous Polyposis Coli Protein
FBXW7	F-box and WD repeat domain containing 7 E3 ubiquitin protein ligase	0.193 (45)	negative regulation of SREBP signaling pathway	0.165	Proto-Oncogene Proteins B-raf
SOX9	SRY (sex determining region Y)-box 9	0.315 (71)	regulation of cell proliferation involved in tissue homeostasis	0.127	Adenomatous Polyposis Coli Protein
TP53	tumor protein p53	0.669 (145)	negative regulation of macromitophagy	0.110	Organoplatinum Compounds
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog	0.675 (147)	GMP binding	0.099	Proto-Oncogene Proteins B-raf
TCF7L2	transcription factor 7-like 2 (T-cell specific HMG-box)	0.736 (160)	beta-catenin-TCF7L2 complex	0.096	Adenomatous Polyposis Coli Protein
MSH6	mutS homolog 6 (E. Coli)	0.769 (166)	oxidized purine DNA binding	0.095	Organoplatinum Compounds
MSH3	mutS homolog 3 (E. Coli)	0.775 (167)	oxidized purine DNA binding	0.095	Organoplatinum Compounds

was breast cancer. Among the highly mutated genes identified by the whole genome and exome sequencing of breast tumors are PIK3CA, TP53, GATA3, CDH1, RB1, MLL3, MAP3K1 and CDKN1, which were previously observed in clinical breast cancer tumors [Ellis et al., 2012]. The

genes not previously observed in those tumors were TBX3, RUNX1, LDLRAP1, STNM2, MYH9, AGTR2, STMN2, SF3B1. Both sets of genes were ranked by the literature-based prioritization algorithm. The top genes from those sets within the Top 1% of the human genome are listed in Table 3.17.

The BRCA1 and BRCA2 are confirmed genetic markers of increased susceptibility to breast

TABLE 3.17: Top Ranking Genes with Respect to the "**Breast Neoplasms**" Phenotype. The Strongest Connection between the Gene Annotation and the Phenotype through the MeSH D Term is Shown for Each Gene. Gene Score and Rank in the Whole Genome with Respect to the "**Breast Neoplasms**" is Shown as Well as Scores of the GO Annotation and the Connecting MeSH D Term

Gene	Description	Gene Score (Gene Rank)	GO Annotation	GO Score	MeSH D
CYP2C19	cytochrome P450 family 2 subfamily C polypeptide 19	0.005 (1)	(R)-limonene 6-monooxygenase activity	0.647	BRCA2 Protein
CYP2C9	cytochrome P450 family 2 subfamily C polypeptide 9	0.011 (2)	(R)-limonene 6-monooxygenase activity	0.647	BRCA2 Protein
BRCA1	breast cancer 1 early onset	0.016 (3)	positive regulation of histone H3-K9 acetylation	0.647	BRCA2 Protein
BRCA2	breast cancer 2 early onset	0.022 (4)	BRCA2-MAGE-D1 complex	0.647	BRCA2 Protein
RB1	Retinoblastoma 1	0.027 (5)	regulation of centromere complex assembly	0.647	BRCA2 Protein
RUNX1	runt-related transcription factor 1	0.525 (112)	positive regulation of progesterone secretion	0.456	Receptors, Progesterone
GATA3	Gata binding protein 3	0.537 (115)	type IV hypersensitivity	0.452	Receptors, Estrogen
TP53	tumor protein p53	0.703 (154)	positive regulation of cell aging	0.417	BRCA1 Protein
CDH1	cadherin 1 type 1 E-cadherin (epithelial)	1.83 (372)	regulation of water loss via skin	0.327	Tamoxifen

cancer and appear at the top of the ranked genome. Since these genes were identified as the top candidates by the developed algorithm it increases confidence about the ability of the prioritization algorithm to identify relevant genes. The other genes from that study: RB1, GATA3, TP53

and CDH1, were placed in the top positions by our algorithm as well. RUNX1 is also among the genes identified as strongly associated with the "**Breast Neoplasms**" phenotype (Table 3.17), although this gene was not confirmed as related to breast neoplasms. The top ranking genes formed the strongest relations to "**BRCA2 protein**" encoded by the BRCA2 gene. Mutations in this gene predispose humans to breast and ovarian cancer. Another chemical that was strongly related to the breast cancer phenotype is "**Tamoxifen**" which is an antineoplastic agent used to treat hormone-sensitive tumors. "**Tamoxifen**" was most strongly linked to the CDH1 (cadherin) gene.

Cancer genes are well characterized and widely studied in literature. Genes, not previously reported as carrying clinically important mutations in studies of colorectal and breast cancer [Ellis et al., 2012, Muzny et al., 2012] had very strong links with the cancer phenotypes in question and were ranked as the top candidates. The ranking already included associations from literature applying indirect inference through chemicals. In colorectal cancer the gene SOX9, although lacking evidence, was linked to the "**Colorectal Neoplasms**" phenotype through the term "**Adenomatous Polyposis Coli Protein**", the same entity which contributed to the genes already confirmed as associated with colorectal cancer. Also, RUNX1, although lacking evidence, was linked to breast cancer through the "**Receptors, progesterone**" term. The lack of expression of progesterone and estrogen receptors and growth factor receptor 2 in breast tumors is known as very aggressive triple negative breast cancer.

3.5 Discussion. Advantages and Limitations of the Literature-Based Prioritization Algorithm

3.5.1 Factors Affecting the Prioritization

The developed prioritization algorithms use literature and gene ontology as main data sources in order to deduce the strongest link between genes and phenotypes. A single data source cannot capture all the relevant information about such links. In addition, the data sources can be biased. MEDLINE is inherently biased towards the most studied topics. Gene Ontology is biased towards better characterized genes. The topics that have received the most attention will have more articles in MEDLINE. If the community focus is towards studying certain genes then it leads to those genes being better annotated. In considering these limitations, genes which were not ranked highly by our algorithm should not be completely ignored as possible candidates. In the current implementation, the proposed literature-based algorithm will not be able to process phenotypes that cannot be adequately expressed in MeSH terms only. In the following sections, factors affecting literature-based prioritization algorithms will be discussed using benchmark examples.

3.5.1.1 Most Studied Topics

The bias arising from more studied topics in literature is illustrated by using the "**Respiration**" phenotype from the benchmark. The test set related to this phenotype consists of 22 genes from the MGI database. Only the Grin1 (Score=0.4873, Rank=120) gene appears within the Top 1% of the ranked whole mouse genome. Following Grin1, the second and third genes that appear within the Top 5% of the ranked genome are Mecp2 and Ache.

The MeSH D terms which have strongest relationships with the "**Respiration**" phenotype are "**Medetomidine**", "**Xylazine**", "**Butorphanol**", "**Nalbuphine**", "**Helium**", "**Oxymorphone**",

"**Anesthetics**" and "**Dissociative**". The connection between "**Oxygen**" and "**Respiration**" is rather weak compared to the other chemicals. "**Oxygen**" is linked to 480 GO annotations but because of the weak connection between "**Oxygen**" and "**Respiration**" these GO annotations do not receive strong weights. The GO annotations formed the strongest associations with the "**Respiration**" phenotype through the MeSH terms representing sedative and analgesic drugs.

The MEDLINE articles annotated by the term "**Respiration**" mostly talk about anesthesia in surgery and veterinary medicine. The GO Annotations most related to the **Respiration** phenotype are associated with it mostly through analgesic drugs that are used in respiration control in anesthesia and surgery. Because of the most studied topics, anesthesia and veterinary medicine, the genes that are in some way related to analgesic drugs will attain the highest ranks with respect to "**Respiration**"; hence the ranking will be biased towards respiration control by drugs. Extensive details on this phenotype are in Appendix A.5.

3.5.1.2 Well Studied Phenotypes and Genes

The "**DNA Repair**" phenotype and genes associated with it are very well characterized. "**DNA Repair**" was the best performing phenotype in the benchmark. The gene list associated with "**DNA Repair**" has 34 genes. 19 genes (55%) from that list appeared within the Top 1% of the ranked genome. The three top candidates are Ogg1 (Score=0.0203, Rank=5), Sirt1 (Score= 0.0528, Rank= 13) and Lig4 (Score=0.0809, Rank=29). The Ogg1 gene has 100 associated publications, and 36 annotations for mouse and human organisms. This gene is involved in DNA repair and base excision REACTOME and KEGG pathways. The Sirt1 gene has 361 associated publications, and 128 annotations for mouse and human organisms including "**regulation of DNA repair**". The Lig4 gene has 75 associated publications and a total of 57 annotations characterizing processes of DNA ligation in repair, recombination and replication.

The top MeSH D terms connected to "**DNA repair**" are: "**Xeroderma Pigmentosum Group**

D Protein" linked by 370 articles, "**Deoxyribonuclease IV Phage T4-Induced**" linked by 318 articles, "**Thymine DNA Glycosylase**" linked by 55 articles, "**MutS DNA Mismatch-Binding Protein**" related by 291 articles and "**Rad52 DNA Repair and Recombination Protein**" related by 255 articles. The strongest connection is "**DNA Glycosylases**" appearing in 1,840 MEDLNE articles together with "**DNA Repair**".

All of these molecular entities are directly involved in DNA modification processes such as cleavage and excision. The "**DNA Repair**" phenotype has 40,730 associated articles in total (as of September 2013). It is extensively studied in literature and linked to very specific bio-molecules. The proposed algorithm will make reliable predictions for candidate genes with respect to the phenotypes having similar properties to the "**DNA Repair**" phenotype. More extensive details about this phenotype are in Appendix A.6.

3.5.1.3 Current State of Knowledge on Genes

Less annotated genes may be dominated by better annotated multi functional genes in prioritization. This effect is illustrated through the example of the "**Autophagy**" phenotype analyzed in the benchmark. The gene list associated with "**Autophagy**" has 18 genes. None of those genes is within the Top 1% of the ranked genome. The genes associated to the autophagy phenotype in the Mouse Genome Informatics database are: Atg16l1, Atg4b, Atg4c, Atg9a, Becn1, Bnip3l, Cisd2, Clcn7, Grid2, Ido1, Lrrk2, Mtap1s, Nhlrc1, Nod2, Oxct1, Pmp22, Pthlh, Ulk1 and Vcp. All had quite low ranks.

The strongest relationships, the MeSH D terms linked to the "**Autophagy**" in MEDLINE comprise "**Class III Phosphatidylinositol 3-Kinases**" (46 articles), "**Lysosomal-Associated Membrane Protein 2**" (51 articles) and "**Ubiquitinated Proteins**" (26 articles). These MeSH D terms designate very broad classes of processes. They are strongly linked to the GO annotations representing many different functions including the functions, directly related to the autophagy

phenotype.

For example, the term "**Class III Phosphatidylinositol 3-Kinases**" has the strongest relationship with "**diaphragm morphogenesis, regulation of G1/S transition checkpoint**" and "**regulation of cytokine secretion involved in the immune response**". The term "**Ubiquitinated Proteins**" has strongest relationships with GO annotations such as "**M Phase, regulation of histone H3-K36 methylation**", "**regulation of transcription from RNA polymerase II promoter by glucose**", "**condensin core heterodimer**", "**branched-chain amino acid catabolic process to alcohol via Ehrlich pathway**" and 443 more annotations that describe very heterogeneous cellular processes.

The phenotypes linked to ubiquitous bio-molecules such as "**Class III Phosphatidylinositol 3-Kinases**" and "**Ubiquitinated Proteins**" will favor multifunctional genes in prioritization. For example, Gata4 appears top first in the ranked whole genome with respect to the "**Autophagy**" phenotype. The Gata4 gene encodes a transcriptional activator. It regulates genes involved in embryogenesis and in myocardial differentiation. The strongest GO annotation for Gata4 which placed that gene at the top is "**diaphragm morphogenesis**" most strongly linked to "**Class III Phosphatidylinositol 3-Kinases**". The Knnc4 gene having 32 annotations appears in the top second position of the ranked whole genome. The highest scoring annotation for this gene is "**intermediate conductance calcium-activated potassium channel activity**" also having the strongest relation to "**Class III Phosphatidylinositol 3-Kinases**".

The autophagy genes have less annotations compared to the top ranking genes: Atg4c has 14 annotations, Atg4b has 13 annotations, Atg161l has 14 annotations and Atg9a has 15 annotations. The top ranking Autophagy gene Atg5 (at the top 5th position in the whole ranked genome), having 25 annotations, was not present in the MGI gene list used in the benchmark. This fact shows that the top-ranking genes truly related to the tested phenotypes might not yet be included in the MGI database, which directly affects the estimation of the performance of the

proposed algorithm in the direction of being underestimated. More details about this phenotype are in Appendix A.7.

3.5.1.4 Good Candidates not on the Test List

The MGI gene lists used to test the performance of the proposed prioritization algorithm contain genes known to be associated with the phenotype by experimental knock-out evidence. Some genes truly related to the phenotype of interest might not be included in the test list as it was already noted in the previous Section 3.5.1.3. The "**Vasoconstriction**" phenotype is an example of this situation. The gene list associated with the "**Vasoconstriction**" phenotype has 11 genes, out of which 6 (66%) are placed within the Top 1% of the ranked genome by our algorithm. Other genes appear within the Top 5% of the ranked genome except Kcnmb1 potassium large conductance calcium-activated channel subfamily M beta member 1. The top ranking genes are Agtr2 (Score=0.0528, Rank=17), Cav1 (Score=0.199, Rank=63) and Uts2 (Score=0.0203, Rank=64).

The highest scoring GO annotations have the strongest links to "**Vasoconstriction Agents**" that are drugs used to cause constriction of the blood vessels. Other bio-molecules most strongly connected to "**Vasoconstriction**" are "**Adrenergic alpha-1 Receptor Agonists, Receptor**", "**Endothelin B, Receptors**", "**Purinergic P2X1**" and "**Endothelins**".

The top scoring GO annotations which are"**vasoconstriction of artery involved in baroreceptor response to lowering of systemic arterial blood pressure**", "**regulation of systemic arterial blood pressure by norepinephrine-epinephrine**" and also 332 other annotations of the top ranking benchmark genes acquired their weights from "**Vasoconstriction Agents**".

The same term "**Vasoconstriction Agents**" contributed most to the weights of the the top ranking whole genome genes not included in the benchmark gene list such as Agt (angiotensinogen (serpin peptidase inhibitor clade A member 8)) and Ace (angiotensin I converting enzyme (peptidyl-dipeptidase A) 1). Evidence from the literature indicates that the Agt gene regulates

vasoconstrictor effects [Szekeres et al., 2012] and the Ace gene influences renovascular hypertension [Ceroni et al., 2010]. The endothelin genes Edn1 and Edn2 also ranked as the top genes. The Edn1 gene is known to be involved in vasoconstriction [Kurihara et al., 1995]. This example illustrates again that some genes known to be related to the phenotype are not yet present in the MGI database. More information on this phenotype is in Appendix A.8.

3.5.2 Advantages

The proposed algorithm has important advantages compared to the other gene prioritization algorithms, in addition to G2D, reviewed in the Introduction.

3.5.2.1 Gene-Phenotype Relationships are Inferred without Reading the Full Text of the Articles

Due to the incorporation of MeSH terms, inferred relationships between phenotypes and Gene Ontology annotations do not depend on the explicit mentions of these relationships in articles. One does not need to process the full texts of scientific articles to establish the links.

The membership function m(phenotype, chemical) in Figure 3.9 corresponding to Equation 3.8 links two MeSH terms by the fuzzy membership value proportional to the number of articles annotated by both MeSH terms. Large membership values indicate strong connections between terms. These connections are meaningful since the MeSH terms describe the content of those articles. For example, if most of the articles about "**Cell Fusion**" are also about "**Viral Fusion Proteins**" then one can conclude that a meaningful relationship exists between these two concepts. This is established without examining the text of the articles (however, to determine the nature of the connections made the full text of the articles should be read).

3.5.2.2 Gene-Phenotype Links Can Be Interpreted in Natural Language

Due to the use of the inclusion relationship, formed gene-phenotype links can be explained in terms of natural language.

The inclusion relationship models semantic linkages between concepts in such a way that narrower concepts are included in broader concepts (outlined in Section 3.2.2). It is directional and better suited to inferring the links between phenotype concepts and functional gene annotations via the intermediate MeSH D terms of molecular entities. The strength of the connection between a phenotype and GO annotation is estimated by applying operations on fuzzy binary relationships that can be interpreted in natural language. In other words, the proposed algorithm links the well-defined meanings of MeSH terms and Gene Ontology annotations to each other by means of modelling a semantic inclusion relationship between the terms. This semantic relationship allows one to interpret formed links between the unique terms in the pair (phenotype, GO annotation) as natural language statements of the following form:

if
    the phenotype includes the chemical to some degree m(phenotype, chemical)
and
    the chemical includes the GO annotation to some degree m(phenotype, GO annotation)
then
    the phenotype includes the GO annotation to some degree m(phenotype, GO annotation).

Each GO annotation as illustrated in Figure 3.9 is related to the phenotype to some degree through intermediate links. Each intermediate link assigns a fuzzy membership value to the GO annotation. Therefore, each GO annotation is characterized by a vector of the values of the fuzzy membership functions shown in square brackets next to the GO annotation. The maximum values indicate the strongest connections. Each intermediate link is characterized by its own fuzzy membership value and should be interpreted separately.

Queries in the PolySearch tool are formulated as linguistic statements. The GENIA [Thompson et al., 2011] corpus contains structured sentence information which is utilized for semantic

interpretation of biomedical texts.

3.5.2.3 Specific and Rare Relationships are Promoted

The inclusion relationship between the terms utilized in the proposed algorithm promotes rare and specific relationships between the chemicals and gene ontology annotations.

The fuzzy binary inclusion relationship (Section 3.2.2) semantically models the phenotype as a broader concept than the chemicals related to it, and models the chemicals as broader concepts that include functional gene annotations. The membership function m(chemical, GO annotation) shown in Figure 3.9 defined by Equation 3.10 connects MeSH D terms (drugs and chemicals) to functional gene annotations by using information in the *gene2pubmed* and *gene2go* NCBI databases. The strength of the connection between unique terms in the pair (MeSH D terms, GO annotation) is modeled as inclusion FBR defined in Equation 3.10. The strength of the connection is computed by counting how many genes share the same MeSH D term and GO annotation, normalized by the number of genes present in *gene2go* that are annotated by that specific GO annotation. Such a definition of membership function promotes rare and specific connections between chemicals and functional gene annotations, giving them higher scores.

For example, suppose that there is only one article which studied a specific single gene and this article is annotated by a specific MeSH D term representing the domain in which the gene was investigated. The studied gene has a very specific function annotated by a GO annotation not shared by any other gene. Such a situation will be reflected by the fuzzy membership function in Equation 3.10 by assigning the highest membership value to the pair (MeSH D term, GO annotation) equal to 1. General GO annotations such as protein binding, shared by many genes, will have weak connections with MeSH D terms. Weak connections are not informative since they imply that although the interaction between the chemical and the gene function may exist, it is not specific because this function is shared by many genes. On the other hand, strong

connections are informative since they imply that there might be strong interactions between the studied drug or chemical and that specific gene function. The approach implemented in the proposed algorithm ensures that uncommon specific relationships between chemicals and functional gene annotations will dominate in the inference of links between the phenotype and functional annotations of the genes. Subsequently, these uncommon relationships will impact the ranking of the genes by placing those carrying the unique functions most strongly related to the phenotype of interest at the top of the ranked list. This implementation of inference is unique to the proposed algorithm.

3.5.2.4 Physical and Chemical Domains Related to Gene Functions Are Identified

Molecular entities linking the phenotype and Gene Ontology annotations represent important background information which reflects current knowledge in the literature.

Functional gene annotations are linked to the phenotype by multiple intermediate connections through different molecular entities (illustrated by chemicals in Figure 3.9). These molecular entities identify the physical and chemical domains in which genes have been studied. MEDLINE topics that cover these molecular entities provide background information in areas of interest in which the studied genes and their functions had more or less impact. These areas present a coarse summary of the current state of knowledge about specific domains through which specific gene functions affecting phenotypes of interest can be inferred.

The existing prioritization systems PosMed, PolySearch and G2D have the ability to inspect background topics. Similarly, the general purpose system Anni2.0 [Jelier et al., 2008] allows the creation of textual profiles to explore the contexts related to various biomedical concepts. However, greater availability of this background knowledge is instrumental to identifying under-represented areas of gene research. For example, molecular entities that have weaker links with phenotypes may indicate under-represented or otherwise interesting areas that are worth

researching, even though it is possible that such weak links are formed by chance. Information about which chemicals are linked to both the phenotype and functional annotations of genes may help uncover unexpected roles of genes in phenotypes that have not been sufficiently explored in the past. This feature of the proposed algorithm provides a broad view of the physical domains bridging functional gene annotations with phenotypes.

3.5.3 Limitations

The discussed computations underlying the proposed algorithm are prone to inherent biases incurred by (i) the topics most studied in MEDLINE, (ii) the most studied genes being better annotated and (iii) the heterogeneity of the functional annotations of genes that play different roles in multiple biological processes.

3.5.3.1 Topics Most Studied in MEDLINE

Rising and strong research interests towards certain topics are reflected by the increased number of MEDLINE articles devoted to those topics.

This phenomenon affects the strength of the links between phenotypes and chemicals computed by the fuzzy membership function m(phenotype, chemical) in Figure 3.9. Topics of greater interest defined by the MeSH terms in MEDLINE will form stronger and larger numbers of connections with chemicals. If a phenotype defined by a specific MeSH term happened to be a topic researched in a different context than defined in the MeSH vocabulary, then the intended meaning of that phenotype definition may be masked. The formed connections between terms will be more a result of the context of phenotype discussion in articles rather than of the phenotype itself.

An example of such bias is illustrated by an analysis of the "Respiration" phenotype in Section 3.5.1.1. One would expect that "**Respiration**" should be strongly connected to "**Oxygen**".

However, MeSH D terms of sedative and analgesic drugs dominate in the strongest connections to the "**Respiration**" phenotype. This effect is a result of the high number of articles annotated with "**Respiration**", which discuss it in the context of anesthesiology in veterinary medicine. Consequently, the functional gene annotations that determine gene ranking are most strongly linked to sedative and analgesic drugs. For example, the annotation "**bradykinin biosynthetic process**", representing processes and chemical reactions involved in the formation of the peptide hormone bradykinin which causes enlargement of the blood vessels, modulating blood pressure, is the highest scoring GO annotation. This causes the Klk1b1 gene, related to hypertension, to be the first top ranking gene with respect to the "**Respiration**" phenotype. The "**bradykinin biosynthetic process**" annotation is linked to the "**Respiration**" phenotype through the "**Xylazine**" chemical, which is an adrenergic alpha-2 agonist used as a sedative, analgesic and centrally-acting muscle relaxant in veterinary medicine. Blood pressure monitoring is an important part of surgery and is affected by the drugs used in anesthesia. Because the highest ranking genes with respect to the "**Respiration**" phenotype were those whose functional annotations had strongest connections to analgesic and sedative drugs, they may be biased towards functional interactions and involvement in the processes induced by those drugs. It is not currently possible, and this is an area of future research, to automatically control for this bias. To control for this bias manually, users may remove links.

3.5.3.2 Greater-Studied Genes Being Better Annotated

The most studied and better-annotated genes may cause biases in inferring the strongest connections between the terms computed by m(chemical, GO annotation) in Figure 3.9. Better-studied genes will have more annotations and more associated articles. This may mask or diminish connections to other genes that are significantly related to the phenotype. This bias is illustrated by the "**Autophagy**" phenotype example in 3.5.1.3. In this example, the genes that have more

annotations compared to the true autophagy genes dominate the Top 1% of the whole ranked genome. Therefore, averaging the fuzzy membership values for each GO annotation is not appropriate in this approach since doing so diffuses information about which terms represent the strongest connections. A way to control for this bias in gene ranking is to use smaller weights for the genes having significantly more annotations than other candidates in the list.

3.5.3.3 Heterogeneity of Gene Annotations in Genes with Different Roles

Genes that have many different roles in various processes will have numerous annotations reflecting all of their functions, and will be strongly linked to numerous phenotypes.

A GO annotation with a strong link to a specific phenotype might be shared by genes that are both essential to the phenotype expression and involved in many processes that are remotely related to and not essential to that expression. In the proposed algorithm, genes are ranked by the maximum fuzzy membership value of their annotations. Because of this, essential and non-essential genes may have ties. The ties are resolved by comparing the second most-highly scoring annotation in addition to the first. Those genes that are not essential but contain a broad spectrum of the annotations may be placed at the top of the list, pushing the more specific genes towards the lower ranks.

The Gata4 gene serves as an example. This gene encodes the transcription factor that regulates genes involved in embryogenesis and myocardial differentiation and function. Gata4 is linked to many chemicals and phenotypes. In the "**Autophagy**" example in Section 3.5.1.3, Gata4 is the top-ranking gene. There is no substantial evidence that Gata4 might be strongly related to "**Autophagy**", but it may be the top ranked gene because of its ubiquitous nature. Its relation to the "**Autophagy**" phenotype is mediated by the chemical "**Class III Phosphatidylinositol 3-Kinases**", which is involved in vesicular transport and indirectly regulates a wide range of cell-growth processes. The "**Class III Phosphatidylinositol 3-Kinases**" bio-molecules have a

very broad spectrum of biological activity. This term is linked to many GO annotations that represent very heterogeneous gene functions, including the annotations of the Gata4 gene. An a priori weighting of the annotations of multi functional genes may control for this bias.

3.6 Conclusion

The developed algorithm uses two independent approaches to gene prioritization: by the MeSH term and by the training genes. The first approach makes an inference based on fuzzy binary relationships to link genes to phenotypes through intermediate concepts defined by MeSH D terms and gene ontology annotations. This prioritization scheme relies on the current state of knowledge accumulated in MEDLINE articles and Gene Ontology databases to date (as of September, 2013). The second approach uses the definition of a phenotype by genes known to be associated with that phenotype. In this approach, the candidate genes are ranked based on the similarity of their annotations to the annotations of the known genes. Similarity is estimated as the maximum information content of the closest common ancestor of the annotations of the training and candidate genes. This approach relies on similarities in the functional content of the genes. To test the proposed algorithm, a benchmark was created consisting of 110 phenotypes with which lists of the genes known to be associated to those phenotypes were retrieved from the MGI data base. The proposed algorithm performed well in the benchmark, making significant predictions in 87% of all tested gene lists. The limitations of the proposed algorithm consist of biases incurred by the most studied topics in the literature and the most researched genes. These biases can be partially accounted for by examining the evidence on gene-phenotype links formed by the algorithm.

Chapter 4

ChIP-Seq Data Analysis of BCL11B and TAL1 Transcription Factors

This chapter outlines two applications of the developed tools to analyze ChIP-Seq data. The first application concerns an analysis of the transcription factor BCL11B, studied in the laboratory of Dr. M. Brand. The second application demonstrates the identification of over-represented composite motifs in peaks bound by the TAL1 transcription factor with the *preferred distances tool* proposed in this thesis. Each of these applications will be introduced in turn.

BCL11B is involved in T-cell development function, lineage commitment and identity maintenance [Li et al., 2010, Liu et al., 2010b]. BCL11B is important to the survival and function of mature T-cells [Wakabayashi et al., 2003]. The relationship between the BCL11B and TAL1 transcription factors in T-ALL (acute lymphoblastic leukemia) is studied in the laboratory of Dr. M. Brand. It was hypothesized that TAL1 might mediate T-ALL, at least partially acting on genes that BCL11B targets; however, the BCL11B target genes in the T-cell lineage are mostly unknown. In order to analyze BCL11B binding genome-wide, chromatin immunoprecipitation (ChIP) followed by deep sequencing (ChIPseq) was performed in Jurkat T-ALL cells. ChIP was performed using an antibody against BCL11B and normal IgG as a negative control according

to the protocol previously described [Palii et al., 2011]. The data was collected at the Dr. M. Brand laboratory at the Ottawa Hospital Research Institute.

We demonstrated the application of the developed tools to characterize the genome-wide BCL11B binding profile and target genes in the Jurkat cells. Pipeline programs were used to perform quality filtering, alignment to the reference genome, visualization in the UCSC genome browser, identification of BCL11B binding sites and motifs analysis. These steps are discussed in Section 4.1. Lastly, in Appendix Section C.27 an example is presented of the analysis of functional enrichment of the genomic regions of BCL11B by GREAT software [McLean et al., 2010] and how the developed prioritization algorithm can be used to prioritize genes associated with BCL11B binding regions.

The TAL1 transcription factor, studied in the laboratory of Dr. M. Brand, was identified as a key mediator of the vascular repair function of primary human endothelial colony-forming cells (ECFCs) [Palii et al., 2014]. Genome-wide analysis of TAL1 in ECFCs was investigated with chromatin immunoprecipitation coupled with massively parallel DNA sequencing (ChIP-Seq). Pipeline programs were used to perform quality filtering, alignment to the reference genome, trimming and remapping of the reads, visualization in the UCSC genome browser and identification of TAL1 binding sites. The applied computational pipeline led to the identification of 1,679 TAL1 genome-wide binding regions [Palii et al., 2014]. *de-novo* motifs in the DNA sequences of the TAL1 binding regions were identified by using motifRG package [Fong et al., 2012]. Analysis of composite motifs was applied to the TAL1 binding regions to identify whether there are any preferred distances at which the binding sites of TAL1 and other transcription factors tend to significantly co-occur in TAL1 genome-wide binding regions in ECFCs. This application will be discussed in more detail in Section 4.2.

4.1 Pipeline Tools Applied to BCL11B ChIP-Seq data

The BCL11B ChIP-Seq analysis steps represent the general workflow of ChIP-Seq data analysis for the identification of narrow transcription factor binding sites. BCL11B ChIP-Seq data were tested for quality and mapped to the Human reference genome hg19 build [UCSC-Genome-Bioinformatics, 2009] by the MAQ program. The choice of the MAQ aligner was motivated by the need to use the same tools for the analysis of the BCL11B TF as was used in the TAL1 study [Palii et al., 2011]. The results of the quality analysis and alignment are summarized in Table 4.1 in terms of the reads produced at each pipeline step.

TABLE 4.1: Counts of Short Reads Across the Processing Stages

Procedure	**Signal Reads**	**Control Reads**
Total reads	110 093 770	104 742 436
Reads remaining after quality control	107 580 225	102 777 284
Reads remaining after removal of the identical reads, linkers and adaptors	36 983 357	26 311 668
Mapped reads in the MAQ alignment	29 796 616	18 398 513
Uniquely mapped reads used for peak calling	26 219 850	16 217 785

Duplication of short reads in the BCL11B sequencing data was observed and the duplicated reads were removed, reducing the data to approximately 26 million reads available for the identification of BCL11B TFBS. The total remaining reads were sufficient for the subsequent analysis as explained later. In the genome-wide TAL1 transcription factor binding study, TAL1 has a narrow binding site like BCL11B, and 11 million reads [Palii et al., 2011] were used for the analysis. The BCL11B binding ChIP-Seq profile, obtained in the alignment step, was visualized in the UCSC browser. Figure 4.1 shows BCL11B binding in Jurkat cells in close proximity of the known BCL11B target genes RUNX3 and ZBTB7B [Kastner et al., 2010]. The displayed tracks comprise the aligned ChIP-Seq signal. The identified peaks are represented by the track following the track representing the binding profile. The RUNX3 gene has three associated BCL11B peaks

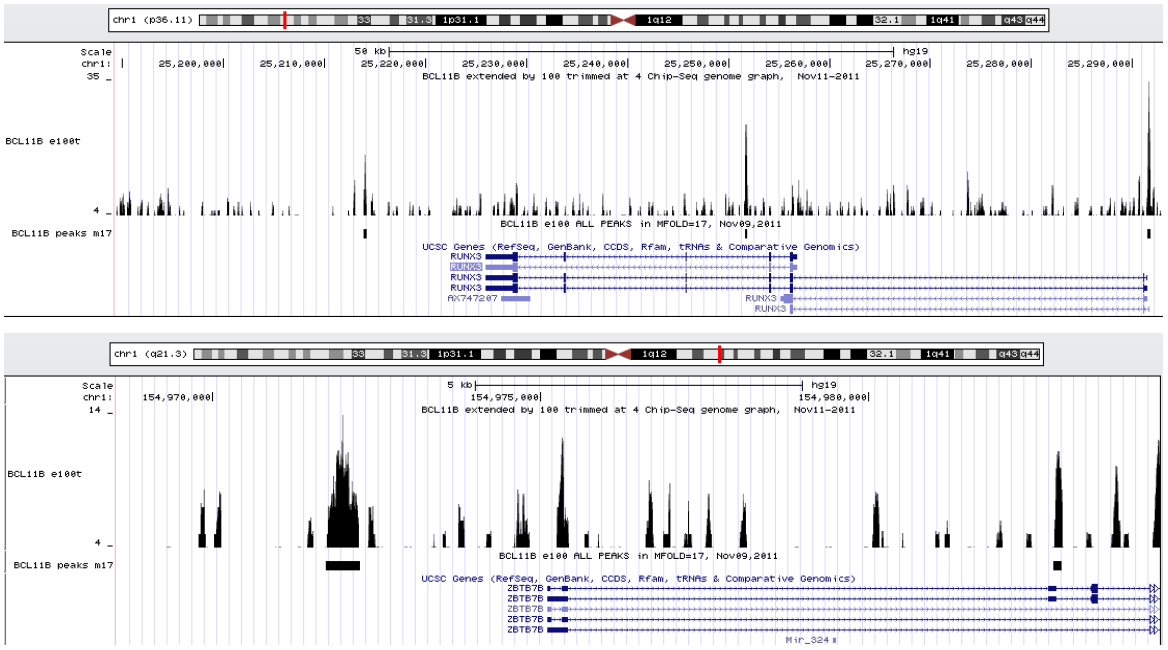


FIGURE 4.1: Visualization of BCL11B Binding with Respect to RUNX3 (Top Panel) and ZBTB7B (Bottom Panel) Genes in the UCSC Genome Browser

and the ZBTB7B gene has two associated BCL11B peaks. 1759 significantly enriched BCL11B binding regions were identified by the MACS program [Zhang et al., 2008] with the following parameters: read length 50bp, sheared DNA fragment length 180bp, threshold for the p value 10^{-7}. The fold of the signal enrichment versus the background was set to 17. The number of identified BCL11B peaks (1759) is in close agreement with the reported 1410 BCL11B binding regions in striatal cells of the mouse brain [Tang et al., 2011].

We tested whether the uniquely mapped reads sufficiently cover the peaks that we wanted to identify (see Section 2.1.1.4). Figure 4.2 shows the rate of conversion of the background reads into the signal as the number of reads used for peak calling increases. Starting with 80% of the total reads, the rate of conversion of the background into the signal becomes small and stable as is seen in Figure 4.2. By increasing the number of sequenced reads no more information is gained to detect new peaks. This test suggests that the number of reads sufficiently covers the genomic

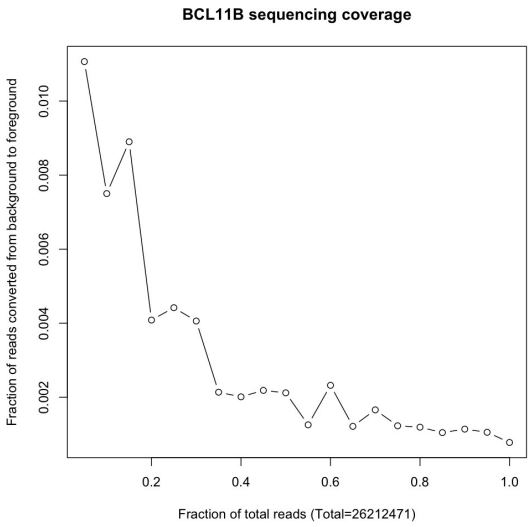


FIGURE 4.2: Rate of Conversion of the Background Reads into the Signal for BCL11B Data.

features (binding regions of the BCL11B transcription factor) which we aimed to identify.

It is important to characterize the BCL11B binding regions in terms of genomic features: promoters, $3'$ and $5'$ untranslated regions, exons, introns and intergenic areas that occur within the binding regions. Such characterization provides information about the possible regulatory properties of the transcription factor. Table 4.2 shows how genomic features are distributed across the BCL11B peaks. A considerable fraction of the peaks are located in the promoter, intron and intergenic areas, similar to the TAL1 TF [Palii et al., 2011]. More peaks identified in the intergenic areas suggest that the BCL11B transcription factor may bind to distal regulatory regions. The peaks in areas with gene promoters concentrate closely around the gene transcription start sites, as illustrated in Figure 4.3. The distances between the peaks and the TSS of the closest genes are estimated by the number of base pairs between the peak summit and the gene transcription start site. The peaks that are upstream of the TSS have negative distances and the peaks downstream have positive distances. The y-axis in Figure 4.3 represents peak frequencies.

Gene-peak association is performed using a criterion of distance in base pairs between a peak

TABLE 4.2: Characterization of the BCL11B Peaks by Occupied Genomic Features

Promoter	**3'Utr**	**5'Utr**	**Exon**	**Intron**	**Intergenic Area**
397	54	180	47	516	565
22.57%	3.07%	10.23%	2.67%	29.33%	32.12%

and a gene TSS. 1513 BCL11B peaks were associated with 1291 genes that have the BCL11B peak within 50Kb up and downstream of their TSS in Jurkat cells. In the developed pipeline, one can set various distance criteria for gene-peak association. In this way one can identify subsets of the possible target genes of the transcription factor that are closer or more distant to the TF binding site. The roles of *some* transcription factors in specific biological processes or cell lines have been characterized [Dunham et al., 2012]. However, the roles and regulatory programs of many TFs are not yet known. One of the strategies to start elucidating these roles is to determine the known target genes of TFs reported by other researchers. These target genes are further verified experimentally.

Some BCL11B target genes were reported in a study of striatal cells in the mouse brain [Kastner et al., 2010]. Intersecting the list of the BCL11B-associated genes in Jurkat cells with the list of genes regulated by BCL11B in murine striatal cells results in 93 common genes listed in Appendix Section C.25. The RUNX3 gene (associated with three BCL11B peaks) and the ZBTB7B gene (associated with one BCL11B peak) from that list were confirmed by RT-qPCR in the laboratory of Dr. M. Brand.

Analysis of transcription factor binding motifs in the identified TF binding sites helps to characterize the properties of the studied TF in terms of its binding specificity and recruitment of other co-factors. TF motifs enriched in peaks can be discovered by *de novo* motif searches. Alternatively, the presence of the motif of interest in the peaks can be quantified by exact sequence matching in the peaks. The *de novo* motif search is performed in peak sequences with respect to the background. Depending on the chosen background the identified motifs may vary. As required, *de novo* identified motifs may be experimentally confirmed by gel shift assays.

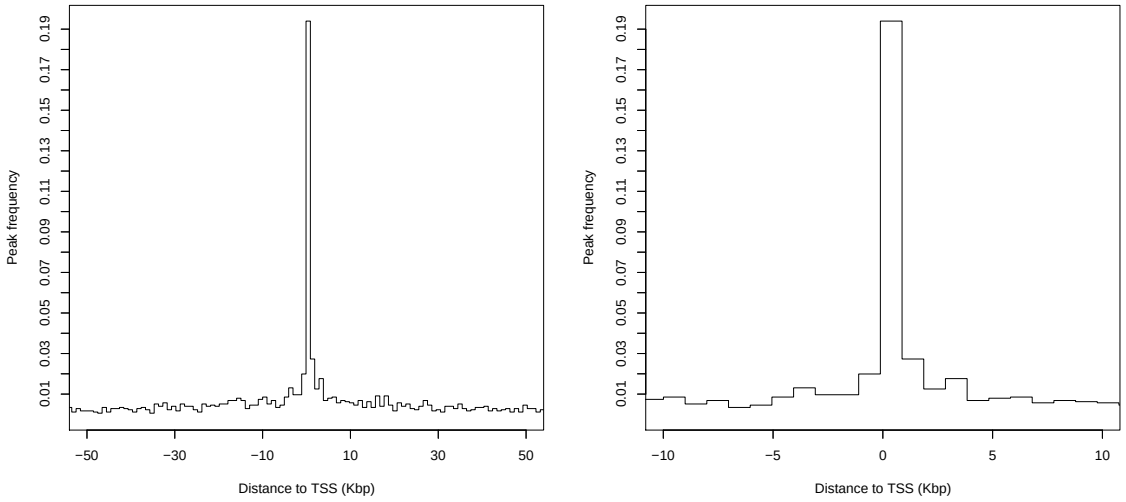


FIGURE 4.3: Peak Frequency at the Distances Within the Region of 50 Kb from the Gene Transcription Start Sites. Distances are Between the BCL11B Peak Summits and the Gene TSS. The Right Panel Represents an Enlarged Area of the Panel on the Left

The pipeline tools allow efficient identification of TF binding regions containing specific sequence motifs and exploration of lists of associated genes. The *in vitro* preferred binding site identified for BCL11B is **5'-GGCCGGAGG-3'** [Avram et al., 2002]. The *in vitro* preferred binding site of BCL11B was present in 55 out of 1759 peaks. The genes associated with the BCL11B peaks having the **5'-GGCCGGAGG-3'** motif are listed in Appendix Section C.26. The gene **ADCY7** from that list is also present in the list of BCL11B targets in murine striatal cells. The most significantly enriched motifs under BCL11B peaks are the ETS motif **5'-SAGGVAG-3'** and the GC-rich motif **5'-DGGCTG-3'** resembling the *in vitro* preferred binding site. The total occurrences of the GC-rich motif sequences are listed in Table 4.3. The GC-rich motif **5'-DGGCTG-3'** was assumed to be the binding motif preferred by BCL11B. In the laboratory of Dr. M. Brand it was hypothesized that TAL1 and BCL11B might interact in the Jurkat cells. This hypothesis can be explored computationally by analyzing composite motifs formed by the TAL1 and BCL11B transcription factors. The motifs associated with the TAL1 transcription factors are EBOX (**5'-CANNTG-3'**) and ETS (**5'-SAGGVAG-3'**)

TABLE 4.3: The Total Occurrences of the **5'-DGGCTG-3'** Motif in 1513 BCL11B Peaks

Pattern	**Total Occurrences**	**Fraction**
GGGCTG	1868	0.42
AGGCTG	1370	0.30
TGGCTG	1277	0.28

[Palii et al., 2011]. A computational analysis was carried out to explore potential evidence for the co-localization of TAL1 binding motifs and the preferred BCL11B binding motif in the Jurkat cells. By using the pipeline tools the BCL11B peaks were scanned for occurrences of the EBOX and ETS motifs in them. All of the BCL11B peaks contained the EBOX motif. The composite motif **EBOX-BCL11B** was found in 1267 BCL11B peaks. The ETS motif was also abundantly present under the BCL11B peaks and the composite motif **ETS-BCL11B** was found in 550 BCL11B peaks. We explored further whether there are preferred distances between these composite motifs in the BCL11B peaks in the Jurkat cells compared to the background in Chromosome 1. Based on this computational analysis none of those composite motifs had any clear pattern of binding preference. The result for **EBOX-BCL11B** is shown in Figure 4.4. Compared to the background there is no substantial evidence that the assumed BCL11B-binding motif **5'-DGCCTG-3'** has a distance preference in co-localizing with EBOX motifs. A specific composite juxtaposed Ets/E-box motif variant, in which ETS is represented by the **5'-CAGGAA(A/G)-3'** sequence, was reported as having a preferred binding distance under TAL1 peaks in Jurkat cells [Palii et al., 2011]. This pattern was computationally explored in the BCL11B peaks. There were 312 BCL11B peaks containing the composite **EBOX-ETS** motif. This composite motif under the BCL11B peaks demonstrated the preferred distance as the composite juxtaposed Ets/E-box motif specific to Jurkat cells in the TAL1 peaks. This is also illustrated in Figure 4.5. The pipeline tools applied to the BCL11B ChIP-Seq data provided the most basic preliminary analysis and characterization of the BCL11B transcription factor. The

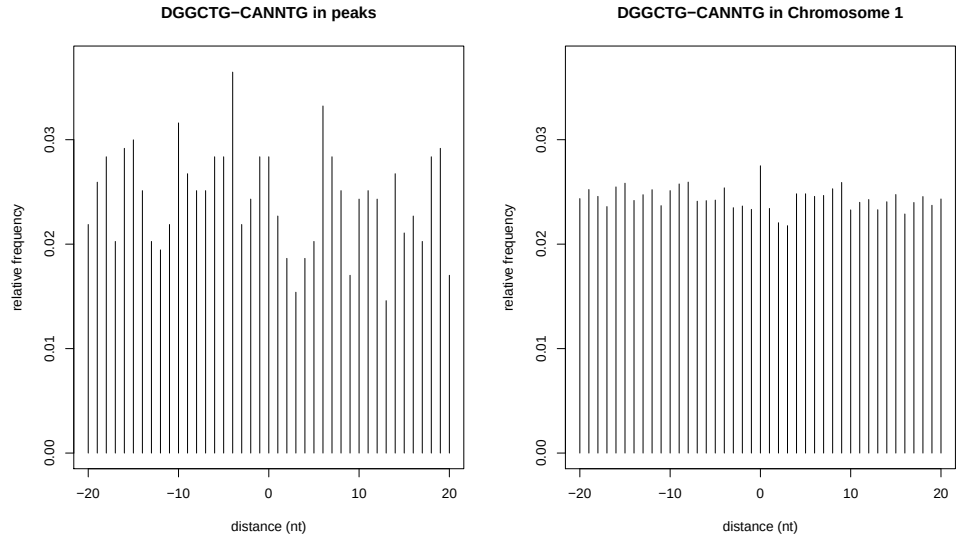


FIGURE 4.4: Frequencies of the Distances in Composite Motifs EBOX-BCL11B in BCL11B Peaks

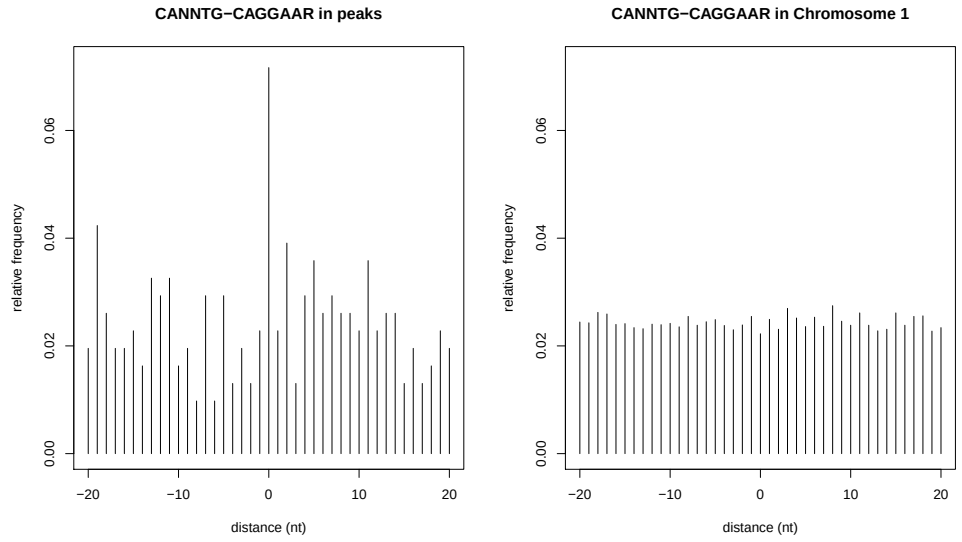


FIGURE 4.5: Frequencies of the Distances in Composite Motifs EBOX-ETS in BCL11B Peaks

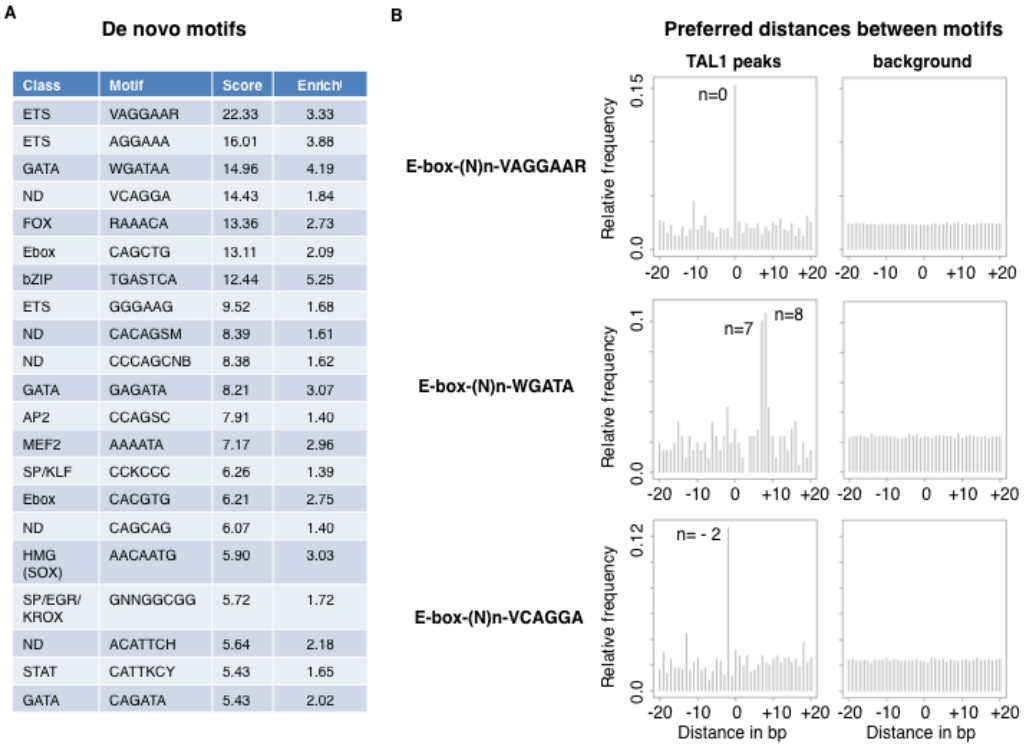
pipeline tools generate data tables containing all pertinent information on peak-gene associations, occurrences of single/composite motifs and intersections of genomic regions. The tables can be filtered and selected subsets can be processed further for extracting more specific information. The pipeline environment ensures the storing of intermediate data that can be retrieved and reanalyzed later without much effort.

4.2 Preferred Distances of TF Co-location in TAL1 Genome-Wide Binding Regions in ECF Cells

The utility of the newly created preferred distances tool is demonstrated in the identification of preferred distances in TAL1 genome-wide binding regions in ECF cells [Palii et al., 2014]. 1,679 TAL1 peaks were identified in the analysis of the ChIP-Seq data of TAL1 genome-wide binding in human endothelial colony-forming cells. DNA sequences that were statistically over-represented in the TAL1 peaks included E-boxes (i.e. TAL1-binding site), and binding sites for the GATA, ETS and FOX families of transcription factors. A complete list of identified motifs and plots showing the preferred distances of TF co-location in peaks is shown in Figure 4.6 (this figure is a reproduction of Figure S4 from the reference [Palii et al., 2014]).

A search for preferred distances identified **EBOX-GATA** and **ETS-EBOX** composite motifs by measuring the relative frequency of preferred distances (in bp) between EBOX and the indicated DNA motif in panel A.

TAL1 transcriptional networks revealed by ChIP-Seq and gene expression profiling in [Palii et al., 2014] suggests that TAL1 is part of a larger network of transcription factors that regulate ECFCs. The motifs that are recognized by transcription factors of the GATA, ETS and FOX families identified in TAL1 peaks suggest that TAL1 could cooperate with these factors to regulate gene expression. The specific composite motifs over-represented in TAL1 peaks in both the ECFCs and the hematopoietic lineage are **EBOX-GATA** and **ETS-EBOX** [Palii et al., 2011]. Other composite motifs seem to be specific to ECF cells, such as a composite with the FOX motif. This suggests that in different biological contexts, TAL1 partners with distinct factors to perform its functions as in ECFCs versus hematopoietic cells. These results highlight the importance of the novel created tool to identify the preferred distances between motifs in the developed ChIP-Seq data analysis pipeline.



Class	Motif	Score	Enrich
ETS	VAGGAAR	22.33	3.33
ETS	AGGAAA	16.01	3.88
GATA	WGATAA	14.96	4.19
ND	VCAGGA	14.43	1.84
FOX	RAAACA	13.36	2.73
Ebox	CAGCTG	13.11	2.09
bZIP	TGASTCA	12.44	5.25
ETS	GGGAAG	9.52	1.68
ND	CACAGSM	8.39	1.61
ND	CCCAGCNB	8.38	1.62
GATA	GAGATA	8.21	3.07
AP2	CCAGSC	7.91	1.40
MEF2	AAAATA	7.17	2.96
SP/KLF	CCKCCC	6.26	1.39
Ebox	CACGTG	6.21	2.75
ND	CAGCAG	6.07	1.40
HMG (SOX)	AACAATG	5.90	3.03
SP/EGR/ KROX	GNNGGCGG	5.72	1.72
ND	ACATTCH	5.64	2.18
STAT	CATTKCY	5.43	1.65
GATA	CAGATA	5.43	2.02

FIGURE 4.6: De-Novo Identified Motifs (Left Panel A) and Plots Showing the Preferred Distances in the Composite Motifs in TAL1 Peaks Identified in ECFCs (Right Panel B).

4.3 Conclusions

The ChIP-Seq data for the transcription factor BCL11B obtained in the Jurkat cells were used to explore the potential target genes of BCL11B. The programs in the ChIP-Seq analysis pipeline were applied to process the BCL11B data. There were 1759 identified BCL11B binding regions/peaks. In the basic analysis, a genome-wide BCL11B binding profile was characterized. BCL11B has a similar binding profile to TAL1 [Palii et al., 2011]. The 1291 genes associated to the BCL11B peaks contain 93 genes which are known BCL11B targets in the striatal mouse brain [Kastner et al., 2010]. 55 BCL11B peaks contained the sequence of the *in vitro* preferred BCL11B binding site.

The full bioinformatics analysis applied to the BCL11B ChIP-Seq data demonstrated all

steps of the analysis that can be applied by using the tools developed in this thesis. These steps were quality filtering, alignment to the reference genome, visualization in the UCSC genome browser, identification of the BCL11B binding sites, analysis of the motifs and analysis of the composite sequences present in peaks. Functional enrichment of the genomic regions of BCL11B by GREAT software and examples of the prioritization of the genes associated with BCL11B binding regions are not the elements of the pipeline. They are presented as supplementary material to this Chapter (in Section C.27). The purpose of including this analysis is to demonstrate how the prioritization algorithm can be applied to a gene list obtained through a ChIP-Seq experiment. The prioritization produced meaningful results "re-discovering" the NOTCH1 gene as strongly linked to the **T Cell Leukemia** Phenotype.

The utility of the novel preferred distances tool in identifying the possible partners of the TAL1 transcription factor in human endothelial colony-forming cells was demonstrated. The tool may become very helpful in revealing the distinct transcriptional networks in which transcription factors participate in various biological contexts. This was illustrated by the different composite motifs of TAL1 with other transcription factors identified in ECFCs and hematopoietic cells.

Chapter 5

General Discussion

This thesis described a newly created gene prioritization algorithm and how tools for composite sequence analysis were integrated into an optimized computational pipeline to analyse ChIP-Seq data. Each of the elements in this process will be discussed in turn.

The proposed prioritization algorithm described here is novel because it is especially useful in formulating biological hypotheses in situations in which little is known about the phenotype and the genes in question, in contrast with existing algorithms that cannot do this. The algorithm begins by linking large lists of poorly characterized genes to non-disease phenotypes described by meaningful keywords through the MEDLINE database. These lists may originate from a variety of sources, including microarray experiments, ChIP-Seq experiments identifying transcription factors' target genes, and scientific literature. The algorithm then deduces which of the links between the genes and phenotypes are strongest and presents the results in an organized manner. This is significantly different from existing algorithms in terms of the methods used to define phenotypes of interest and infer their relationships with genes. As of now, biologists still find it challenging to interpret large lists of poorly characterized genes with these algorithms, which are somewhat limited in terms of how they define phenotypes. To better understand how the new algorithm is unique, the existing algorithms were parsed into three categories and examined.

5.1 Overview of Findings

The first category of existing algorithms only uses known phenotype-related genes, or training genes, while the second focuses solely on human disease phenotypes. The third category uses general keywords from literature to define phenotypes. All must deduce how genes and phenotypes are related by mining selected information sources, retrieving and integrate data from them, but there are some differences between them. Each will be discussed in turn.

Algorithms based on the use of training genes known to be related to the phenotype of interest, as in the Endeavour [Tranchevent et al., 2008] and ToppGene [Chen et al., 2009] tools, make prioritizations on the basis of pattern classification [De Bie et al., 2007]. Training genes extract phenotype-defining information from various data sources based on how similar the phenotype is to the training genes, and then build a model of a phenotype based on this extraction [Tranchevent, 2011]. In other words, the model represents gene features that are most characteristic of that specific phenotype. The candidate genes are ranked by how similar their features are to the features of the model. For example, Endeavour relies on genomic data fusion from multiple information sources [Aerts et al., 2006]. This tool may be very useful if the properties of the training genes clearly define the phenotype properties of interest in the organisms being investigated. Knowing these properties, one can characterize candidate genes by comparing them to the training genes. Genes that have similar characteristics to the training genes may also play an important role in previously unknown phenotype expressions. This principle of discovery is known as "guilt by association" [Wang and Marcotte, 2010].

Although very useful in detecting similarities between candidates and training genes, the integration of data from multiple sources has limitations. The main limitation is in existing schemes of combining the information from different sources to rank the candidates [Aerts et al., 2006]. First, the prioritization algorithms using training genes generally differ with respect to the data sources they use [Tranchevent et al., 2011]. Different information sources of training

genes lead to different models and similarity metrics. Second, some data sources do not have complete information on some genes, so if the phenotype in question has not been sufficiently studied and there are no genes known to be associated with it, then the training genes approach is not effective. Third, the training genes might represent a heterogeneous group biasing phenotype definition in some way. For example, the data fusion scheme relies on the independence of information sources about gene properties, but in practice they are not entirely independent. Protein-protein interaction databases, the gene interaction databases and gene ontology refer to scientific publications as supporting evidence for the information they store, and might even be derived from the literature. While it should do so, the scheme does not account for these possible interactions and overlaps between sources. To remedy this, Endeavour attempted to use the one class support vector machine with radial function kernels [De Bie et al., 2007, Tranchevent, 2011] for data fusion. Each data source was modeled by three kernels. The classifier was trained to find the best separating hyperplanes encapsulating training genes. The trained classifiers were used to classify candidates. However, this approach was investigated theoretically without clearly showing its advantages.

Many tools specific to human diseases use phenotype definitions from databases [Perez-Iratxeta et al., 2005, van Driel et al., 2006]. Because human disease phenotypes have been extensively studied and are well represented by OMIM [Hamosh et al., 2005], and because they contain structured information suited to uncovering meaningful links between human diseases and genes, it is relatively easy to associate genes with said phenotypes. However, phenotypes other than diseases and phenotypes in Mammalian Phenotype Ontology [Kohler et al., 2009] are not yet represented by well-structured and information-rich resources [Groth et al., 2010, Houle et al., 2010].

The third category of algorithms, which use general keywords from literature to define phenotypes, are exemplified by tools such as PosMed, PolySearch, GeneProspector and CANDID

[Cheng et al., 2008, Hutz et al., 2008, Yoshida et al., 2009, Yu et al., 2008b]. These tools rely on finding matching documents in MEDLINE or locally-created databases, and then associating genes with the matching documents. General purpose discovery-oriented systems such as iHOP [Fernandez et al., 2007], Anni2.0 [Jelier et al., 2008], Arrowsmith [Smalheiser and Swanson, 1998, Smalheiser et al., 2009] and PosMed [Yoshida et al., 2009], use conceptual networks. Users can browse through the network and create textual profiles describing genes, proteins, or other biomedical concepts. However, once again, there will be genes and processes that are not well represented in literature and there is little information about them that can be retrieved.

Thus, while the obvious advantage offered by specialized gene information databases is that specific information can be extracted very quickly, complementing the literature with more information sources for gene prioritization is advantageous in allowing potential use of algorithms that offer novel interpretations of existing information. The only existing method which provides underlying ideas for the new algorithmic approach proposed here is G2D [Perez-Iratxeta et al., 2002a, 2005], which prioritizes genes with respect to human disease phenotypes. However, the scope of the application of the prioritization algorithm developed here is distinct from G2D. In fact, our algorithm extends G2D by contributing the following:

- The proposed algorithm establishes meaningful links between genes and phenotypes, and enables prioritization, beyond human disease phenotypes by using concepts of MeSH vocabulary from the categories A, C, D, and G, summarized in Table 3.1.

- The conceptual universe of the proposed prioritization method consists of MeSH terms and Gene Ontology annotations. In contrast, G2D defines only disease phenotypes through OMIM [Hamosh et al., 2005] identifiers and maps the definitions into the associated multiple MeSH terms of the diseases.

- The proposed algorithm can be applied beyond human organisms as the annotated genes of the entire genome for human (Homo sapiens), mouse (Mus musculus) and fly (Drosphila melanogaster) are prioritized. In contrast, G2D focuses on human genes.

- The data in the *gene2go* and *gene2pubmed* NCBI databases [Maglott et al., 2011] are used to link GO annotations to MeSH terms of Drugs and Chemicals describing the molecular entities. In contrast, G2D works with the RefSeq database for this purpose [Perez-Iratxeta et al., 2002a].

- The proposed algorithm utilizes fuzzy binary inclusion relationships between concepts, based on mathematical operations of fuzzy set theory [Zimmermann, 2010, 1996], to infer gene-phenotype links. Binary links between genes and concepts encoded by MeSH terms are computed as values of fuzzy membership functions (see Section 3.2.3). In contrast, G2D uses a similarity relationship [Perez-Iratxeta et al., 2002a].

The proposed algorithm is an attempt to remedy some of the challenges presented by information shortages and the way existing algorithms described above are configured to define phenotypes and determine relationships. The proposed algorithm has important advantages compared to the other gene prioritization algorithms, in addition to G2D, reviewed extensively in Section 3.5.2.

5.2 Examining Evidence of Gene-Phenotype Relationships

The majority of the existing gene prioritization tools reviewed above have limited abilities to explore the reasons for phenotype-gene association, with the exception of PosMed and G2D. But using the proposed algorithm, evidence supporting the obtained gene rankings can be easily examined. As a result of inference, the MeSH D terms which are most strongly related to both the candidate genes and phenotype are identified. This is useful as it reveals the physical background domains related to the candidate genes without the need to process or read the text

of all associated articles. The availability of this background information opens up the possibility of identifying and examining unique aspects of the functions of the studied genes.

However, a single information source cannot account for all aspects of gene relations to phenotypes even if MeSH vocabulary contains information about the processes, phenomena and phenotypes studied in literature. And while functional gene annotations are also associated with scientific publications, there will be genes and processes that are not well represented in literature, as stated earlier. In this situation inferring links between genes and phenotypes might be more effective using other information sources, as genes can also be characterized by their interactions with other molecular entities, by their sequences and by the information about the protein domains of the products. These gene properties can be retrieved computationally from specialized databases.

The proposed gene prioritization algorithm uses three information sources to establish links between the phenotypes and the genes: literature, Gene Ontology databases and homology between the genes of the human, mouse and fly. If genes known to be related to the phenotype are available then they can also be utilized for phenotype definition. For a phenotype defined by known genes, the prioritization of candidates is based on estimating a semantic similarity between the Gene Ontology annotations of the known genes and the candidates. This prioritization principle is entirely different from the literature-based prioritization discussed earlier, as it takes into account only functional similarities between genes. The utilization of known genes for phenotype definition in the proposed algorithm compensates for the lack of possible evidence in the literature for the connection between a gene and phenotype if that gene has not been extensively studied.

This method of prioritization via estimated functional similarity between known genes and candidates adds a new line of evidence to support gene-phenotype connections. The genes that

have not been thoroughly studied, or have fewer annotations, might be ranked low by literature-based algorithms. However, if at least a few genes known to be related to that phenotype can be used to augment the phenotype definition, then the candidate genes "missed" by a literature-based algorithm will be "noticed" by the algorithm using the known genes to define the phenotype. Just a few annotations of the candidate that are semantically similar to the annotations of the training genes are enough for the true candidate to be identified by the algorithm.

5.2.1 Application Testing

Section 3.4.1 discussed the application of the proposed algorithm to prioritize genes related to the "**Cell Fusion**" phenotype. The prioritization algorithm was applied to characterize the roles of genes induced by the signaling of MAPK-p38, whose expression is required for myoblast fusion into multinucleated myotubes [Liu et al., 2012]. The gene Cd53 from the p38 dependant gene list was newly identified using the STRING data base [Szklarczyk et al., 2011] as a potential fusogen based on its ability to interact with proteins known to be mediating myoblast fusion; this was experimentally confirmed [Liu et al., 2012]. The proposed algorithm also successfully identified Cd53 as the candidate involved in cell fusion processes. The p38 dependant gene list of phenotypes closely related to cell fusion was prioritized. A summary of the phenotypes and the prioritization results are presented in Table 3.14. The definitions of the phenotypes were augmented by two genes, Cd81 and Cd9, known to be implicated in cell fusion processes [Dittmar and Zanker, 2011]. Regarding the "**Membrane Fusion**" phenotype, both the literature-based prioritization algorithm and the algorithm based on known genes independently assigned high ranks to the same genes of a small group pulled from the p38 dependent gene list. Cd53 was among the genes in this group.

The Cd53 gene has not been extensively studied and had only 8 gene ontology annotations:

"**protein binding**", "**positive regulation of myoblast fusion**", "**signal transduction**", "**cell surface**", "**cell-cell junction**", "**immunological synapse**", "**integral to membrane**" and "**plasma membrane**", and is not well represented in literature. This gene was not ranked highly by the literature-based algorithm with respect to other cell fusion-related phenotypes, but was ranked highly with respect to the specific membrane fusion phenotype defined by both the MeSH term and the known genes. In this particular example, the phenotype defined by the known genes Cd81 and Cd9 can be interpreted as a positive control for the phenotype "**Membrane Fusion**", since it is known that those two genes affect fusion processes on the level of cellular membranes.

Another example concerns the use of the putative myoblast cell fusion genes to augment the "**Cell Fusion**" phenotype. In the list of the genes induced by the signaling of MAPK-p38, the Myh9 gene, known to be related to the cell fusion, was identified as a good candidate through the training genes line of evidence. Through the same line of evidence the Edn1 and Fb1 genes were also pointed out by the algorithm as strong candidates for the cell fusion phenotype. The Edn1 gene was identified as a strong candidate also by the literature-based line of evidence. These examples show the potential of augmenting phenotype definition with known genes in identifying the relevant candidates. The algorithm using the phenotype definition as a keyword and the algorithm using the phenotype defined by known genes are based on entirely different algorithmic principles, although both of them use GO annotations to rank genes. The former algorithm links the concepts by exploiting the MEDLINE and NCBI gene databases; the latter is solely based on the functional similarity between the genes. Yet the results of both algorithms agree well.

5.3 Evaluation of Performance

To estimate the performance of text mining algorithms three core elements are required: the task of the algorithm, the gold standard and evaluation metrics [Shatkay and Craven, 2012].

The task defines what the algorithm is supposed to do. The gold standard represents a benchmark collection of examples with known answers to test the algorithm. The evaluation metrics represent objective functions to estimate the algorithm's performance.

In terms of the standard, a benchmark of 110 phenotypes with associated gene sets was created. MGI genes [Bult et al., 2013], which are associated with the phenotypes defined in Mammalian Phenotype Ontology [Smith and Eppig, 2012] by knock out evidence, were used. A recall of the gene prioritization algorithm in the benchmark gene sets was used as an evaluation metric. The recall of the proposed algorithm for each phenotype was estimated as a percentage of known genes ranked within the Top 1% and 5% of the whole genome.

The median size of the benchmark gene sets was 32. The median recall within the Top 1% of the ranked genome was 23.53%, and within the Top 5% it was 51.69%. 87% of the tested gene sets had significant valid predictions within the Top 1% of the ranked genome. Based on the benchmark estimates, on average we had $\sim$ 7, 8 significant predictions within the Top 1% and 16, 17 within the Top 5% of the whole genome for an average size gene set of 32 genes. The performance of the proposed algorithm in the benchmark conforms to the standard performances of prioritization algorithms [Bornigen et al., 2012, Moreau and Tranchevent, 2012], according to the criteria used by other benchmarks of gene prioritization algorithms (outlined in Section 3.3). It has been suggested in other studies [Bornigen et al., 2012, Moreau and Tranchevent, 2012] that well-performing gene prioritization algorithms are expected to rank a small subset of the test genes within the Top 5%-15% of the whole genome. The performance of the proposed algorithm was estimated within the Top 1% of the whole genome, which is five times more stringent than the standard Top 5% criterion [Bornigen et al., 2012].

The usual strategy for comparing prioritization tools is to estimate how they rank the disease gene with respect to randomly selected candidates. For example, a study which compared prioritization algorithms used 42 genes known to be associated with diseases [Bornigen et al.,

2012, Yu et al., 2008a] and the algorithms ranked them with respect to randomly selected candidates. Endeavour and ToppGene were reported as the best in ranking the known disease genes highly. These algorithms were assessed in a cross-validation procedure [Bornigen et al., 2012, Yu et al., 2008a], which involved removing one disease test gene from the training genes, including it in a long list of random candidates and then prioritizing the random candidates. There were many runs for each disease test gene. The average rank of each disease test gene was computed across all runs. If the prioritization ranked those genes within the top 5%-15% of the ranked random candidate list, then it was assumed that it performed well. If prioritization is being assessed with respect to ranking the whole genome, then the criterion of the top 5%-15% ranked genome genes containing a small subset of the test genes is also suggested as indicator of good performance [Moreau and Tranchevent, 2012]. However, this criterion is rather arbitrary since there is no known report in the literature arguing in favor of this threshold by formal statistical or mathematical reasoning.

We tried to reproduce the prioritization results for known disease genes stated in the references [Bornigen et al., 2012, Moreau and Tranchevent, 2012]. We used phenotypes whose keywords had adequate matches to the MeSH disease terms. More detailed summary of the prioritization results obtained by the algorithm developed in this thesis is presented in Appendix D.27. Our results compare very favourably with the well performing methods reported in these references [Bodenreider and Burgun, 2010, Moreau and Tranchevent, 2012] such as Endevour, Pinta and SUSPECTS.

It was already noted that benchmarking gene prioritization algorithms is challenging [Moreau and Tranchevent, 2012]. This occurs for several reasons. First, there is a lack of established criteria for what constitutes a good performance. Second, it is difficult to collect substantial experimental evidence to show that a gene is definitively linked to some phenotype of interest; conversely, it is difficult to prove that the gene is not related to the phenotype in any way. To

put it concretely, the random list of candidates which usually serves as a background to test the performance of a tool might not be entirely random. Suppose we have test gene A, known to be related to a cellular phenotype, and a set of random gene candidates which will be used to evaluate a rank of the test gene. It is possible to have a gene B among the random candidates which is related to the cellular phenotype in ways previously unknown. If the random gene B receives a higher rank than gene A, then the tool which is being compared was able to capture this relationship because of its algorithm. Thus, the ability to assess the performance of prioritization algorithms in a highly informed way is always limited. Third, there is a lack of publicly-available benchmark sets of gene-phenotype relationships. The experimentally validated gene-phenotype relationships accumulated in the MGI database are naturally based on the current research interests of the groups submitting results to that database. The documented gene sets in MGI available to the public represent small fractions of the mouse genome genes and cover a limited number of specifically investigated phenotypes. Such available gene sets were used to benchmark the proposed gene prioritization algorithm. Therefore, the conclusions drawn about the performance of the proposed algorithm were based only on an inherently limited number of the genes available for testing. Naturally, these limited gene sets do not include other genes that might be strongly related to the tested phenotype of interest.

Even so, the benchmark performance of the algorithm proposed in this thesis might be underestimated because the genes truly related to the phenotypes in the MGI database may not have been documented yet. The possibility of this underestimated performance was illustrated with the example of ranking genes with respect to the "**Vasoconstriction**" phenotype in Section 3.5.1.4. The two top ranking genes Agt and Ace are not present in the MGI benchmark gene set, but they are clearly related to the "**Vasoconstriction**" phenotype, as evidenced by the literature [Ceroni et al., 2010, Szekeres et al., 2012]. If all true top-ranking gene candidates were included in the benchmark gene sets, then the performance estimates of the proposed gene prioritization

algorithm would be much higher. This example supports the claim that the proposed prioritization algorithm has better performance than estimated in the benchmark. Overall, inherent limitations imposed by the available content in the MGI database cannot be alleviated except by ongoing research.

5.4 Integrative Analysis of ChIP-Seq Data

The ChIP-Seq data analysis pipeline was outlined in Chapter 2. The novel aspect of this pipeline is the incorporation of a new tool to analyze composite motif sequences. This tool identifies the distances at which studied transcription factors pretend to co-occur in peaks containing the binding sites of those TFs.

Many possibilities exist in performing complete ChIP-Seq data analysis, starting with the raw data and ending the identification of regions containing transcription factor binding sites. ChIP-Seq data analysis pipelines answer most of the needs of researchers in different ways. Genome-wide ChIP-Seq pipelines are generally composed of input data management, alignment, peak detection, genome-wide characterization of the identified peaks, sequence and motif analysis and association of peaks to genes. The most notable system for ChIP-Seq data using Galaxy as a back end is Cistrome [Liu et al., 2011]. Cistrome, besides peak calling, integrates analysis tools such as gene annotation, motif and pathway analysis. CisGenome [Ji et al., 2008] is a software system for the analysis of tiling ChIP-chip arrays and ChIP-Seq data. CisGenome is distributed as an installable package. It supports peak detection, motif analysis and peak annotation by the nearest genes. Nebula [Boeva et al., 2012] is a Galaxy server mostly addressing alignment and peak detection needs.

With respect to the software tools for analysis, researchers encounter significant challenges in making optimal choices for their purposes [Nekrutenko and Taylor, 2012]. For example, Bioconductor [Gentleman et al., 2005], which requires the R language interpreter, has powerful

tools for microarray and DNA sequence analysis. However, alignment programs such as MAQ [Li et al., 2008], BWA [Li and Durbin, 2009, 2010], Bowtie [Langmead et al., 2009] and the individual peak calling programs such as MACS [Feng et al., 2012, Zhang et al., 2008], SISSR [Jothi et al., 2008] and PeakSeq [Rozowsky et al., 2009], are widely used and better-documented than similar tools in R [Mercier et al., 2011, Morgan et al., 2009]. Integrative data analysis environments such as Galaxy [Blankenberg et al., 2011], Nebula [Boeva et al., 2012] and Cistrome [Liu et al., 2011], have been created to ease the use of available ChIP-Seq analysis tools for biologists without specialized training in bioinformatics. In the same vein, tools packaged into machine images are publicly available in the computing cloud [Krampis et al., 2012] to leverage the power of distributed computing.

But even after making the choices for the most appropriate tools the need for a convenient interface to interact with the programs remains [Halling-Brown and Shepherd, 2008]. There is a distinction between pipelines utilizing command line tools and pipelines that utilize user interfaces to access integrated tools. Each category has its own advantages and limitations. Pipelines utilizing command line tools offer greater flexibility in creating workflows and taking advantage of available computing infrastructure, enabling one to run the programs in parallel. Such pipelines are published as protocols comprising programmatic details and scripts [Bardet et al., 2012]. Users have a high level of control in accommodating such a pipeline for their own computational needs. However, for users lacking bioinformatics training, working with command line pipelines may become very challenging. The pipelines that offer user interfaces do not require specialized knowledge. However, they restrict the level of control over tools that the user may want to have. Neither the command line nor fully integrated pipelines can fully answer all of the needs of users.

Thus, in this thesis a hybrid approach was taken with regard to implementing the ChIP-Seq pipeline. Some functionalities of our pipeline have already been addressed by other packages.

The novel tools of the pipeline consist of the randomization of DNA sequences, the peak-gene association using various rules and the analysis of composite sequences to identify the preferred distances of motif co-location in peaks. The tools employ Chromosome 1 of the human genome build hg19 as a background. Using the pipeline, one can quickly perform a basic ChIP-Seq data analysis to get a result and to generate a hypothesis. In-depth further analysis can be performed with other specialized tools.

ChIP-Seq data analysis components, such as alignment to the reference genome and peak calling, are already standard routine, with the peak calling being less standardized. However, algorithms for the analysis of transcription factor motifs in regulatory regions are still being actively developed [Aerts et al., 2008], and although the importance of analyzing composite sequences and preferences of transcription factor binding in various biological contexts is already recognized [Kulakovskiy and Makeev, 2013, Kulakovskiy et al., 2011, Palii et al., 2011], there is no publicly available software for that purpose. It has been argued that the majority of predicted motif binding sites might be non-functional [Wasserman and Sandelin, 2004]. The presence of composite sequences of several motifs binding at characteristic distances in peaks renders those peaks more likely to contain functional binding sites. Two distinct motifs occurring at the preferred distance in a set of peaks is less likely to happen just by chance if this preference is not observed in the genomic background.

These are all reasons why the pipeline leverages cluster computing infrastructure to execute computationally demanding tasks. The computationally-demanding and time-consuming tasks of alignment to the reference genome and peak calling, requiring a high level of control over the parameters, were implemented through the command line. The tools for downstream analysis of the peaks and DNA sequence motifs were integrated and made accessible through the Galaxy workflow management system interface. This integration of the computational tools into the

ChIP-Seq pipeline with composite sequence analysis generates complete fast results for the ChIP-Seq data starting with the original short reads and ending with genome-wide characterization of transcription factor binding. This is one of the main tasks of downstream peak analysis [Giannopoulou and Elemento, 2011, Ji et al., 2008]; others include the identification of motifs in the peaks [Bailey, 2011, Bailey et al., 2009, Heinz et al., 2010] and summarizing the identified motifs.

Complete characterization consists of the following:

- classification of peak regions into the categories of promoter, 5' and 3' untranslated regions, exons, introns and inter-genic areas based on whether the peak intersects those gene features;
- association of peaks to genes;
- motif-matching and motif-discovery;
- analysis of the composite sequence in peaks;
- making the plots.

The existing integrated peak analysis pipeline [Giannopoulou and Elemento, 2011] provides a tool for the analysis of composite sequences by creating comprehensive summaries of identified known motifs co-locating with each other. However, it does not offer the functionality to explore the composite sequences defined by a user in the peaks and the genomic background. The pipeline implemented in this thesis addresses this shortcoming by integrating a new tool which allows the user to generate composite motif sequences and match them to the peaks and the background. This tool was implemented in R using its package for the manipulation of strings, stringr [Wickham, 2012] and the Biostrings package [Morgan et al., 2009] in Bioconductor. By means of this tool the user can identify whether two motifs prefer to co-locate in peaks at

some distance. In the composite sequence the two motifs are separated by a certain number of bases. The composite sequences of the motifs separated by various distances are created. Each composite is matched to the peak sequences through exact pattern-matching and the counting of all matches. If the composite with the motifs at a certain distance is enriched in peaks, then this distance is hypothesized to be a preferred distance at which the studied motifs co-occur in the peaks. The pattern of the preferred distances in peaks is compared to the pattern in the background sequences. The preferred distance is not expected to be seen in the background. The presence of the preferred distance in peaks and its absence in the background indicate that the two motifs tend to co-locate at a characteristic distance in the biological context of the ChIP-Seq experiment.

This tool was verified by identifying the preferred distances of motif co-location in BCL11B and TAL1 peaks in Chapter 4. A use case of the analysis of genome-wide binding of the BCL11B and TAL1 transcription factors and their potential target genes was presented. Within the computational framework developed in this thesis, ChIP-Seq data of the transcription factors were analyzed. Their peaks were characterized in terms of: (i) genomic features (promoter, intron, exon and other) occurring in peaks, (ii) genes closest to peaks and (iii) motifs present in peaks and their composite sequences.

The main advantage of the approach taken in this thesis consists of the management of tools for ChIP-Seq data analysis by utilizing a modular structure of Galaxy. The advantages of the hybrid approach and the integration explained above consist of easy exchange and reuse of intermediate data between different analysis workflow tools and a simple user interface on which to inspect results and rerun modified analysis. Moreover, the newly created tools can be easily integrated with existing tools. Advanced features of other pipelines that consist of separate modules can be easily plugged into the Galaxy workflow management system utilized here.

5.5 Future Directions

The objectives and aims stated for this thesis have been achieved. However, the gene prioritization algorithm developed in this thesis could be improved with added functionality. The incorporation of more data sources, such as protein-protein interactions and shared protein domains, would add more lines of evidence on gene-phenotype relationships. Similar principles of fuzzy inference, as in linking MeSH D terms with GO annotations, can be applied to retrieve information from those added data sources. Using the protein-protein interaction data, the gene ontology annotations of the interacting partner genes could be added to the GO annotations of each gene, thus expanding functional gene descriptions to embrace the functional domains of their interacting partners. Similarly, for shared protein domain data, the GO annotations of the proteins with which a gene product shares a domain can be added to the functional description of the gene. These simple improvements would greatly increase the power of the proposed algorithm to detect gene-phenotype relationships with more confidence. Further, it would be beneficial to incorporate the data from genome-wide association studies into our framework. The relationships between the genomic regions and phenotypes elucidated in the GWAS can be incorporated as a separate fixed information table. The descriptions of the GWAS phenotypes can be mapped to the MeSH terms. The associated genomic regions can be mapped to the closest genes. The lines of evidence containing GWAS data would represent fixed, already-known relationships. This line of evidence could also be used as a partial ground truth to support other relationships derived by the proposed algorithm.

The gene prioritization tool developed here uses a database containing relationships between MeSH terms. Information in this database can be used in other applications, such as to characterize gene lists by mapping the topics that are strongly related to the genes in the list. Such analysis would be similar to the enrichment of annotations, but in this case the annotations represent MeSH terms. The enriched GO annotations in the analyzed gene list would show the

MeSH D terms most strongly connected to those GO annotations. The connected MeSH D terms would identify the domains of chemical and physical phenomena in which the genes from the list are most studied and also would link a user to the most relevant literature. Such functionality may be useful for database curators. An application programming interface supporting various queries to the gene prioritization database would be another useful improvement. Communicating with this database through the application programming interface would make it more accessible; this could be easily integrated into the Galaxy workflow management system, thereby providing extended functionality for the new ChIP-Seq data analysis.

5.6 Conclusion

The work presented in this thesis attempted to answer the current need for algorithms that can be used in the analysis of composite motif sequences in ChIP-Seq data. It also attempted to address the need for a gene prioritization algorithm to analyze gene-phenotype relationships in which non-disease phenotypes are defined by meaningful keywords. The presented gene prioritization algorithm is meant to be used to infer gene-phenotype links by using the semantic inclusion relationship in the framework of fuzzy set theory. Defining phenotypes with genes known to be related to them can be used to augment phenotype definitions via meaningful keywords. Combined, these two approaches to defining phenotypes offer a powerful tool to derive gene-phenotype links. The importance and novelty of the thesis contribution consists of the following: (i) The created gene prioritization algorithm addresses the lack of tools to analyze gene-phenotype relationships in which non-disease phenotypes are defined by meaningful keywords. Outputs of this algorithm provide starting points for browsing the evidence supporting the gene-phenotype links established within the current state of knowledge in literature. (ii) The newly created tool for the analysis of the composite motifs in peaks fills a gap in the availability of tools to analyze preferred binding distances in transcription factor regulatory complexes.

(iii) The implemented ChIP-Seq data analysis computational pipeline produces expedited results that can be shared and efficiently reused by the integrated tools during different stages of analysis. It provides an easy way to manage various computational tools implemented in different programming languages. The implemented gene prioritization algorithm provides an instrument to generate and explore plausible biological hypotheses. The created gene prioritization tool has been implemented as a web application. Links to gene prioritization application and ChIP-Seq data analysis pipeline tools are available through `http://figshare.com/PhenotypeLinks` project.

References

E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous, and B. S. Pickard. SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, 22(6):773–774, Mar 2006. URL http://www.ncbi.nlm.nih.gov/pubmed/16423925.

S. Aerts. Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. *Curr. Top. Dev. Biol.*, 98:121–145, 2012. URL http://www.ncbi.nlm.nih.gov/pubmed/22305161.

S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L. C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau. Gene prioritization through genomic data fusion. *Nat. Biotechnol.*, 24(5):537–544, May 2006. URL http://www.ncbi.nlm.nih.gov/pubmed/16680138.

S. Aerts, M. Haeussler, S. van Vooren, O. L. Griffith, P. Hulpiau, S. J. Jones, S. B. Montgomery, and C. M. Bergman. Text-mining assisted regulatory annotation. *Genome Biol.*, 9(2):R31, 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18271954.

S. Aerts, S. Vilain, S. Hu, L. C. Tranchevent, R. Barriot, J. Yan, Y. Moreau, B. A. Hassan, and X. J. Quan. Integrating computational biology and forward genetics in Drosophila. *PLoS Genet.*, 5(1):e1000351, Jan 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19165344.

F. Al-Shahrour, R. Diaz-Uriarte, and J. Dopazo. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580, Mar 2004. URL `http://www.ncbi.nlm.nih.gov/pubmed/14990455`.

A. Alexa, J. Rahnenfuhrer, and T. Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–1607, Jul 2006. URL `http://www.ncbi.nlm.nih.gov/pubmed/16606683`.

G. Altobelli. Bioinformatics applied to gene transcription regulation. *J. Mol. Endocrinol.*, 49 (2):R51–59, Oct 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22723668`.

C. Andronis, A. Sharma, V. Virvilis, S. Deftereos, and A. Persidis. Literature mining, ontologies and information visualization for drug repurposing. *Brief. Bioinformatics*, 12(4):357–368, Jul 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21712342`.

A. V. Antonov, T. Schmidt, Y. Wang, and H. W. Mewes. ProfCom: a web tool for profiling the complex functionality of gene groups identified from high-throughput data. *Nucleic Acids Res.*, 36(Web Server issue):W347–351, Jul 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18460543`.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25 (1):25–29, May 2000. URL `http://www.ncbi.nlm.nih.gov/pubmed/10802651`.

R. K. Auerbach, B. Chen, and A. J. Butte. Relating genes to function: identifying enriched transcription factors using the ENCODE ChIP-Seq significance tool. *Bioinformatics*, 29(15): 1922–1924, Aug 2013. URL `http://www.ncbi.nlm.nih.gov/pubmed/23732275`.

D. Avram, A. Fields, T. Senawong, A. Topark-Ngarm, and M. Leid. COUP-TF (chicken ovalbumin upstream promoter transcription factor)-interacting protein 1 (CTIP1) is a sequence-specific DNA binding protein. *Biochem. J.*, 368(Pt 2):555–563, Dec 2002. URL `http://www.ncbi.nlm.nih.gov/pubmed/12196208`.

C. Backes, A. Keller, J. Kuentzer, B. Kneissl, N. Comtesse, Y. A. Elnakady, R. Muller, E. Meese, and H. P. Lenhof. GeneTrail–advanced gene set enrichment analysis. *Nucleic Acids Res.*, 35(Web Server issue):W186–192, Jul 2007. URL `http://www.ncbi.nlm.nih.gov/pubmed/17526521`.

T. L. Bailey. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, 27(12):1653–1659, Jun 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21543442`.

T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, 37(Web Server issue):W202–208, Jul 2009. URL `http://www.ncbi.nlm.nih.gov/pubmed/19458158`.

A. F. Bardet, Q. He, J. Zeitlinger, and A. Stark. A computational pipeline for comparative ChIP-seq analyses. *Nat Protoc*, 7(1):45–61, Jan 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22179591`.

A. D. Baxevanis. The importance of biological databases in biological discovery. *Curr Protoc Bioinformatics*, Chapter 1:Unit 1.1, Jun 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21633941`.

K. R. Blahnik, L. Dou, H. O'Geen, T. McPhillips, X. Xu, A. R. Cao, S. Iyengar, C. M. Nicolet, B. Ludascher, I. Korf, and P. J. Farnham. Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. *Nucleic Acids Res.*, 38(3):e13, Jan 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/19906703`.

J. M. Blanca, L. Pascual, P. Ziarsolo, F. Nuez, and J. Canizares. ngs_backbone: a pipeline for read cleaning, mapping and SNP calling using next generation sequence. *BMC Genomics*, 12: 285, 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21635747`.

D. Blankenberg, N. Coraor, G. Von Kuster, J. Taylor, and A. Nekrutenko. Integrating diverse databases into an unified analysis framework: a Galaxy approach. *Database (Oxford)*, 2011: bar011, 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21531983`.

O. Bodenreider and A. Burgun. A framework for comparing phenotype annotations of orthologous genes. *Stud Health Technol Inform*, 160(Pt 2):1309–1313, 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/20841896`.

V. Boeva, A. Lermine, C. Barette, C. Guillouf, and E. Barillot. Nebula–a web-server for advanced ChIP-seq data analysis. *Bioinformatics*, 28(19):2517–2519, Oct 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22829625`.

A. Boorsma, B. C. Foat, D. Vis, F. Klis, and H. J. Bussemaker. T-profiler: scoring the activity of predefined groups of genes using gene expression data. *Nucleic Acids Res.*, 33(Web Server issue):W592–595, Jul 2005. URL `http://www.ncbi.nlm.nih.gov/pubmed/15980543`.

D. Bornigen, L. C. Tranchevent, F. Bonachela-Capdevila, K. Devriendt, B. De Moor, P. De Causmaecker, and Y. Moreau. An unbiased evaluation of gene prioritization tools. *Bioinformatics*, 28(23):3081–3088, Dec 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/23047555`.

A. P. Boyle, L. Song, B. K. Lee, D. London, D. Keefe, E. Birney, V. R. Iyer, G. E. Crawford, and T. S. Furey. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.*, 21(3):456–464, Mar 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21106903`.

M. D. Brazas, D. Yim, W. Yeung, and B. F. Ouellette. A decade of Web Server updates at the Bioinformatics Links Directory: 2003-2012. *Nucleic Acids Res.*, 40(Web Server issue): W3–W12, Jul 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22700703`.

C. J. Bult, J. T. Eppig, J. A. Blake, J. A. Kadin, and et al. The mouse genome database: genotypes, phenotypes, and models of human disease. *Nucleic Acids Res.*, 41(Database issue): D885–891, Jan 2013. URL `http://www.ncbi.nlm.nih.gov/pubmed/23175610`.

A. J. Butte and I. S. Kohane. Creation and implications of a phenome-genome network. *Nat. Biotechnol.*, 24(1):55–62, Jan 2006. URL `http://www.ncbi.nlm.nih.gov/pubmed/16404398`.

J. Cairns, C. Spyrou, R. Stark, M. L. Smith, A. G. Lynch, and S. Tavare. BayesPeak–an R package for analysing ChIP-seq data. *Bioinformatics*, 27(5):713–714, Mar 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21245054`.

E. Capriotti, N. L. Nehrt, M. G. Kann, and Y. Bromberg. Bioinformatics for personal genome interpretation. *Brief. Bioinformatics*, 13(4):495–512, Jul 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22247263`.

P. Carmona-Saez, M. Chagoyen, F. Tirado, J. M. Carazo, and A. Pascual-Montano. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol.*, 8(1):R3, 2007. URL `http://www.ncbi.nlm.nih.gov/pubmed/17204154`.

A. E. Carpenter and D. M. Sabatini. Systematic genome-wide screens of gene function. *Nat. Rev. Genet.*, 5(1):11–22, Jan 2004. URL `http://www.ncbi.nlm.nih.gov/pubmed/14708012`.

A. Ceroni, E. D. Moreira, C. T. Mostarda, G. J. Silva, E. M. Krieger, and M. C. Irigoyen. Ace gene dosage influences the development of renovascular hypertension. *Clin. Exp. Pharmacol. Physiol.*, 37(4):490–495, Apr 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/19930431`.

E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, and A. Ma'ayan. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14:128, 2013. URL `http://www.ncbi.nlm.nih.gov/pubmed/23586463`.

J. Chen, H. Xu, B. J. Aronow, and A. G. Jegga. Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics*, 8:392, 2007. URL `http://www.ncbi.nlm.nih.gov/pubmed/17939863`.

J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.*, 37(Web Server issue):W305–311, Jul 2009. URL `http://www.ncbi.nlm.nih.gov/pubmed/19465376`.

D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju, and D. S. Wishart. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res.*, 36(Web Server issue):399–405, Jul 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18487273`.

P. J. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.*, 38(6):1767–1771, Apr 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/20015970`.

T. De Bie, L. C. Tranchevent, L. M. van Oeffelen, and Y. Moreau. Kernel-based data fusion for gene prioritization. *Bioinformatics*, 23(13):i125–132, Jul 2007. URL `http://www.ncbi.nlm.nih.gov/pubmed/17646288`.

D. de la Iglesia D, M. Garcia-Remesal, G. de la Calle, C. Kulikowskiand F. Sanz, and V. Maojo. The impact of computer science in molecular medicine: enabling high-throughput. *Current topics in medicinal chemistry*, 13(5):526–75, 2013. URL `http://www.ncbi.nlm.nih.gov/pubmed/23548020`.

G. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.*, 4 (5):P3, 2003. URL `http://www.ncbi.nlm.nih.gov/pubmed/12734009`.

T. Dittmar and K. Zanker, editors. *Advances in experimental medicine and biology: Cell Fusion in Health and Disease*. Dordrecht NLD. Springer, 2011.

S. W. Doniger, N. Salomonis, K. D. Dahlquist, K. Vranizan, S. C. Lawlor, and B. R. Conklin. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.*, 4(1):R7, 2003. URL `http://www.ncbi.nlm.nih.gov/pubmed/12540299`.

L. du Plessis, N. Skunca, and C. Dessimoz. The what, where, how and why of gene ontology–a primer for bioinformaticians. *Brief. Bioinformatics*, 12(6):723–735, Nov 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21330331`.

J.T. Dudley and K.J. Karczewski. *Exploring personal genomics*. Oxford University Press, 2013.

I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, and et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22955616`.

M. J. Ellis, L. Ding, D. Shen, J. Luo, and et al. Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature*, 486(7403):353–360, Jun 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22722193`.

J. T. Eppig, J. A. Blake, C. J. Bult, J. A. Kadin, and et al. The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Res.*, 40(Database issue):D881–886, Jan 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22075990`.

K. M. Esvelt and H. H. Wang. Genome-scale engineering for systems and synthetic biology. *Mol. Syst. Biol.*, 9:641, 2013. URL http://www.ncbi.nlm.nih.gov/pubmed/23340847.

S. Falcon and R. Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–258, Jan 2007. URL http://www.bioconductor.org/packages/2.12/bioc/html/GOstats.html.

J. Falgueras, A. J. Lara, N. Fernandez-Pozo, F. R. Canton, G. Perez-Trabado, and M. G. Claros. SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. *BMC Bioinformatics*, 11:38, 2010. URL http://www.ncbi.nlm.nih.gov/pubmed/20089148.

J. Feng, T. Liu, and Y. Zhang. Using MACS to identify peaks from ChIP-Seq data. *Curr Protoc Bioinformatics*, Chapter 2:Unit 2.14, Jun 2011a. URL http://www.ncbi.nlm.nih.gov/pubmed/21633945.

J. Feng, T. Liu, B. Qin, Y. Zhang, and X. S. Liu. Identifying ChIP-seq enrichment using MACS. *Nat Protoc*, 7(9):1728–1740, Sep 2012. URL http://www.ncbi.nlm.nih.gov/pubmed/22936215.

X. Feng, R. Grossman, and L. Stein. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. *BMC Bioinformatics*, 12:139, 2011b. URL http://www.ncbi.nlm.nih.gov/pubmed/21554709.

G. H. Fernald, E. Capriotti, R. Daneshjou, K. J. Karczewski, and R. B. Altman. Bioinformatics challenges for personalized medicine. *Bioinformatics*, 27(13):1741–1748, Jul 2011. URL http://www.ncbi.nlm.nih.gov/pubmed/21596790.

J. M. Fernandez, R. Hoffmann, and A. Valencia. iHOP web services. *Nucleic Acids Res.*, 35(Web Server issue):W21–26, Jul 2007. URL http://www.ncbi.nlm.nih.gov/pubmed/17485473.

X. M. Fernandez-Suarez and M. Y. Galperin. The 2013 Nucleic Acids Research Database Issue and the online molecular biology database collection. *Nucleic Acids Res.*, 41(Database issue): 1–7, Jan 2013. URL `http://www.ncbi.nlm.nih.gov/pubmed/23203983`.

Flybase. Links to the model organism projects at the flybase web portal, 2013. URL `http://flybase.org/static_pages/docs/MODs.html`.

A. P. Fong, Z. Yao, J. W. Zhong, Y. Cao, W. L. Ruzzo, R. C. Gentleman, and S. J. Tapscott. Genetic and epigenetic determinants of neurogenesis and myogenesis. *Dev. Cell*, 22(4):721–735, Apr 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22445365`.

R. Frijters, M. van Vugt, R. Smeets, R. van Schaik, J. de Vlieg, and W. Alkema. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Comput. Biol.*, 6(9), 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/20885778`.

H. Frohlich, N. Speer, A. Poustka, and T. Beissbarth. GOSim–an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinformatics*, 8:166, 2007. URL `http://www.ncbi.nlm.nih.gov/pubmed/17519018`.

F. Fuchs, G. Pau, D. Kranz, O. Sklyar, C. Budjan, S. Steinbrink, T. Horn, A. Pedal, W. Huber, and M. Boutros. Clustering phenotype populations by genome-wide RNAi and multiparametric imaging. *Mol. Syst. Biol.*, 6:370, Jun 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/20531400`.

T. S. Furey. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.*, 13(12):840–852, Dec 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/23090257`.

R. Gentleman, V. Carey, W. Huber, R. Irizarry, and Dudoit S., editors. *Bioinformatics abd computational biology solutions using R and Bioconductor*. Statistics for Biology and Health. Springer, 2005. URL `http://bioconductor.org`.

E. G. Giannopoulou and O. Elemento. An integrated ChIP-seq analysis platform with customizable workflows. *BMC Bioinformatics*, 12:277, 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21736739`.

G. V. Gkoutos, E. C. Green, A. M. Mallon, J. M. Hancock, and D. Davidson. Using ontologies to describe mouse phenotypes. *Genome Biol.*, 6(1):R8, 2005. URL `http://www.ncbi.nlm.nih.gov/pubmed/15642100`.

P. Groth, I. Kalev, I. Kirov, B. Traikov, U. Leser, and B. Weiss. Phenoclustering: online mining of cross-species phenotypes. *Bioinformatics*, 26(15):1924–1925, Aug 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/20562418`.

M. Halling-Brown and A. J. Shepherd. Constructing computational pipelines. *Methods Mol. Biol.*, 453:451–470, 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18712319`.

A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, 33(Database issue):D514–517, Jan 2005. URL `http://www.ncbi.nlm.nih.gov/pubmed/15608251`.

A. Hatem, D. Bozda, A. E. Toland, and U. V. Catalyurek. Benchmarking short sequence mapping tools. *BMC Bioinformatics*, 14:184, 2013. URL `http://www.ncbi.nlm.nih.gov/pubmed/23758764`.

R. D. Hawkins, G. C. Hon, and B. Ren. Next-generation genomics: an integrative approach. *Nat. Rev. Genet.*, 11(7):476–486, Jul 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/20531367`.

S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, 38(4): 576–589, May 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/20513432`.

R. Hoehndorf, M. Dumontier, and G. V. Gkoutos. Evaluation of research in biomedical ontologies. *Brief. Bioinformatics*, Sep 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22962340`.

M. M. Hoffman, J. Ernst, S. P. Wilder, A. Kundaje, R. S. Harris, M. Libbrecht, B. Giardine, P. M. Ellenbogen, J. A. Bilmes, E. Birney, R. C. Hardison, I. Dunham, M. Kellis, and W. S. Noble. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.*, 41(2):827–841, Jan 2013. URL `http://www.ncbi.nlm.nih.gov/pubmed/23221638`.

D. A. Hosack, G. Dennis, B. T. Sherman, H. C. Lane, and R. A. Lempicki. Identifying biological themes within lists of genes with EASE. *Genome Biol.*, 4(10):R70, 2003. URL `http://www.ncbi.nlm.nih.gov/pubmed/14519205`.

D. Houle, D. R. Govindaraju, and S. Omholt. Phenomics: the next challenge. *Nat. Rev. Genet.*, 11(12):855–866, Dec 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/21085204`.

D. Hristovski, B. Peterlin, J. A. Mitchell, and S. M. Humphrey. Using literature-based discovery to identify disease candidate genes. *Int J Med Inform*, 74(2-4):289–298, Mar 2005. URL `http://www.ncbi.nlm.nih.gov/pubmed/15694635`.

d. a. W. Huang, B. T. Sherman, Q. Tan, J. R. Collins, W. G. Alvord, J. Roayaei, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki. The DAVID Gene Functional Classification

Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.*, 8(9):R183, 2007. URL `http://www.ncbi.nlm.nih.gov/pubmed/17784955`.

d. a. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, 37(1): 1–13, Jan 2009. URL `http://www.ncbi.nlm.nih.gov/pubmed/19033363`.

M. Huang, A. Neveol, and Z. Lu. Recommending MeSH terms for annotating biomedical articles. *J Am Med Inform Assoc*, 18(5):660–667, 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21613640`.

J. E. Hutz, A. T. Kraja, H. L. McLeod, and M. A. Province. CANDID: a flexible method for prioritizing candidate genes for complex human traits. *Genet. Epidemiol.*, 32(8):779–790, Dec 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18613097`.

R. Jelier, M. J. Schuemie, A. Veldhoven, L. C. Dorssers, G. Jenster, and J. A. Kors. Anni 2.0: a multipurpose text-mining tool for the life sciences. *Genome Biol.*, 9(6):R96, 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18549479`.

H. Ji, H. Jiang, W. Ma, D. S. Johnson, R. M. Myers, and W. H. Wong. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, 26(11):1293–1300, Nov 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18978777`.

M. Jiang, J. Anderson, J. Gillespie, and M. Mayne. uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*, 9:192, 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18405375`.

R. Jothi, S. Cuddapah, A. Barski, K. Cui, and K. Zhao. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.*, 36(16):5221–5231, Sep 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18684996`.

D. Karolchik, A. S. Hinrichs, and W. J. Kent. The UCSC Genome Browser. *Curr Protoc Bioinformatics*, Chapter 1:Unit1.4, Dec 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/23255150`.

P. Kastner, S. Chan, W. K. Vogel, L. J. Zhang, A. Topark-Ngarm, O. Golonzhka, B. Jost, S. Le Gras, M. K. Gross, and M. Leid. Bcl11b represses a mature T-cell gene expression program in immature CD4(+)CD8(+) thymocytes. *Eur. J. Immunol.*, 40(8):2143–2154, Aug 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/20544728`.

P. Kersey and R. Apweiler. Linking publication, gene and protein data. *Nat. Cell Biol.*, 8(11): 1183–1189, Nov 2006. URL `http://www.ncbi.nlm.nih.gov/pubmed/17060904`.

P. V. Kharchenko, M. Y. Tolstorukov, and P. J. Park. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, 26(12):1351–1359, Dec 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/19029915`.

S. Y. Kim and D. J. Volsky. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6:144, 2005. URL `http://www.ncbi.nlm.nih.gov/pubmed/15941488`.

D. C. Koboldt, L. Ding, E. R. Mardis, and R. K. Wilson. Challenges of sequencing human genomes. *Brief. Bioinformatics*, 11(5):484–498, Sep 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/20519329`.

S. Kohler, S. Bauer, D. Horn, and P. N. Robinson. Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, 82(4):949–958, Apr 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18371930`.

S. Kohler, M. H. Schulz, P. Krawitz, S. Bauer, S. Dolken, C. E. Ott, C. Mundlos, D. Horn, S. Mundlos, and P. N. Robinson. Clinical diagnostics in human genetics with semantic

similarity searches in ontologies. *Am. J. Hum. Genet.*, 85(4):457–464, Oct 2009. URL `http://www.ncbi.nlm.nih.gov/pubmed/19800049`.

S. Kohler, S. C. Doelken, A. Rath, S. Ayme, and P. N. Robinson. Ontological phenotype standards for neurogenetics. *Hum. Mutat.*, 33(9):1333–1339, Sep 2012a. URL `http://www.ncbi.nlm.nih.gov/pubmed/22573485`.

S. Kohler, S. C. Doelken, A. Rath, S. Ayme, and P. N. Robinson. Ontological phenotype standards for neurogenetics. *Hum. Mutat.*, 33(9):1333–1339, Sep 2012b. URL `http://www.ncbi.nlm.nih.gov/pubmed/22573485`.

O. Korol and M. Turcotte. Learning relationships between over-represented motifs in a set of dna sequences. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2012 IEEE Symposium on*, pages 144–149, 2012. doi: 10.1109/CIBCB.2012.6217223.

M. Krallinger, F. Leitner, and A. Valencia. Analysis of biological processes and diseases using text mining approaches. *Methods Mol. Biol.*, 593:341–382, 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/19957157`.

K. Krampis, T. Booth, B. Chapman, B. Tiwari, M. Bicak, D. Field, and K. E. Nelson. Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. *BMC Bioinformatics*, 13:42, 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22429538`.

I. V. Kulakovskiy and V. J. Makeev. DNA Sequence Motif: A Jack of All Trades for ChIP-Seq Data. *Adv Protein Chem Struct Biol*, 91:135–171, 2013. URL `http://www.ncbi.nlm.nih.gov/pubmed/23790213`.

I. V. Kulakovskiy, A. A. Belostotsky, A. S. Kasianov, N. G. Esipova, Y. A. Medvedeva, I. A. Eliseeva, and V. J. Makeev. A deeper look into transcription regulatory code by preferred pair

distance templates for transcription factor binding sites. *Bioinformatics*, 27(19):2621–2624, Oct 2011. URL http://www.ncbi.nlm.nih.gov/pubmed/21852305.

Y. Kurihara, H. Kurihara, H. Oda, K. Maemura, R. Nagai, T. Ishikawa, and Y. Yazaki. Aortic arch malformations and ventricular septal defect in mice deficient in endothelin-1. *J. Clin. Invest.*, 96(1):293–300, Jul 1995. URL http://www.ncbi.nlm.nih.gov/pubmed/7615798.

T. D. Laajala, S. Raghav, S. Tuomela, R. Lahesmaa, T. Aittokallio, and L. L. Elo. A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC Genomics*, 10:618, 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/20017957.

S. G. Landt, G. K. Marinov, A. Kundaje, and et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, 22(9):1813–1831, Sep 2012. URL http://www.ncbi.nlm.nih.gov/pubmed/22955991.

B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3):R25, 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19261174.

B. R. Lee, S. Cho, Y. Song, S. C. Kim, and B. K. Cho. Emerging tools for synthetic genome design. *Mol. Cells*, 35(5):359–370, May 2013. URL http://www.ncbi.nlm.nih.gov/pubmed/23708771.

M. Leleu, G. Lefebvre, and J. Rougemont. Processing and analyzing ChIP-seq data: from short reads to regulatory interactions. *Brief Funct Genomics*, 9(5-6):466–476, Dec 2010. URL http://www.ncbi.nlm.nih.gov/pubmed/20861161.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009. URL `http://www.ncbi.nlm.nih.gov/pubmed/19451168`.

H. Li and R. Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, Mar 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/20080505`.

H. Li and N. Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinformatics*, 11(5):473–483, Sep 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/20460430`.

H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18(11):1851–1858, Nov 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18714091`.

L. Li, M. Leid, and E. V. Rothenberg. An early T cell lineage commitment checkpoint dependent on the transcription factor Bcl11b. *Science*, 329(5987):89–93, Jul 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/20595614`.

R. Li, C. Yu, Y. Li, T. W. Lam, S. M. Yiu, K. Kristiansen, and J. Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, Aug 2009. URL `http://www.ncbi.nlm.nih.gov/pubmed/19497933`.

E. T. Liu, S. Pott, and M. Huss. Q and A: ChIP-seq technologies and the study of gene regulation. *BMC Biol.*, 8:56, 2010a. URL `http://www.ncbi.nlm.nih.gov/pubmed/20529237`.

P. Liu, P. Li, and S. Burke. Critical roles of Bcl11b in T-cell development and maintenance of T-cell identity. *Immunol. Rev.*, 238(1):138–149, Nov 2010b. URL `http://www.ncbi.nlm.nih.gov/pubmed/20969590`.

Q. C. Liu, X. H. Zha, H. Faralli, H. Yin, C. Louis-Jeune, E. Perdiguero, E. Pranckeviciene, P. Munoz-Canoves, M. A. Rudnicki, M. Brand, C. Perez-Iratxeta, and F. J. Dilworth. Comparative expression profiling identifies differential roles for Myogenin and p38 MAPK signaling in myogenesis. *J Mol Cell Biol*, 4(6):386–397, Dec 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22847234`.

T. Liu, J. A. Ortiz, L. Taing, C. A. Meyer, B. Lee, Y. Zhang, H. Shin, S. S. Wong, J. Ma, Y. Lei, U. J. Pape, M. Poidinger, Y. Chen, K. Yeung, M. Brown, Y. Turpaz, and X. S. Liu. Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.*, 12(8):R83, 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21859476`.

Z. Lombard, C. Park, K. D. Makova, and M. Ramsay. A computational approach to candidate gene prioritization for X-linked mental retardation using annotation-based binary filtering and motif-based linear discriminatory analysis. *Biol. Direct*, 6:30, 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21668950`.

K. L. MacQuarrie, Z. Yao, A. P. Fong, S. J. Diede, E. R. Rudzinski, D. S. Hawkins, and S. J. Tapscott. Comparison of genome-wide binding of MyoD in normal human myogenic cells and rhabdomyosarcomas identifies regional and local suppression of promyogenic transcription factors. *Mol. Cell. Biol.*, 33(4):773–784, Feb 2013. URL `http://www.ncbi.nlm.nih.gov/pubmed/23230269`.

S. Maere, K. Heymans, and M. Kuiper. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–3449, Aug 2005. URL `http://www.ncbi.nlm.nih.gov/pubmed/15972284`.

D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, 39(Database issue):D52–57, Jan 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21115458`.

M. Mahner and M. Kary. What exactly are genomes, genotypes and phenotypes? And what about phenomes? *J. Theor. Biol.*, 186(1):55–63, May 1997. URL `http://www.ncbi.nlm.nih.gov/pubmed/9176637`.

S. Mahony and P. V. Benos. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, 35(Web Server issue):W253–258, Jul 2007. URL `http://www.ncbi.nlm.nih.gov/pubmed/17478497`.

A. Manconi, E. Vargiu, G. Armano, and L. Milanesi. Literature retrieval and mining in bioinformatics: state of the art and challenges. *Adv Bioinformatics*, 2012:573846, 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22778730`.

A. J. Marian. Challenges in medical applications of whole exome/genome sequencing discoveries. *Trends Cardiovasc. Med.*, 22(8):219–223, Nov 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22921985`.

A. Masoudi-Nejad, A. Meshkin, B. Haji-Eghrari, and G. Bidkhori. Candidate gene prioritization. *Mol. Genet. Genomics*, 287(9):679–698, Sep 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22893106`.

C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, and G. Bejerano. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, 28(5):495–501, May 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/20436461`.

E. Mercier, A. Droit, L. Li, G. Robertson, X. Zhang, and R. Gottardo. An integrated pipeline for the genome-wide analysis of transcription factor binding sites from ChIP-Seq. *PLoS ONE*, 6(2):e16432, 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21358819`.

L. R. Meyer, A. S. Zweig, A. S. Hinrichs, D. Karolchik, R. M. Kuhn, M. Wong, C. A. Sloan, K. R. Rosenbloom, G. Roe, B. Rhead, B. J. Raney, A. Pohl, V. S. Malladi, C. H. Li, B. T. Lee, K. Learned, V. Kirkup, F. Hsu, S. Heitner, R. A. Harte, M. Haeussler, L. Guruvadoo, M. Goldman, B. M. Giardine, P. A. Fujita, T. R. Dreszer, M. Diekhans, M. S. Cline, H. Clawson, G. P. Barber, D. Haussler, and W. J. Kent. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.*, 41(Database issue):D64–69, Jan 2013.

S. Miyamoto. Information retrieval based on fuzzy associations. *Fuzzy sets and systems*, 38, 1990.

Y. Moreau and L. C. Tranchevent. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet.*, 13(8):523–536, Aug 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22751426`.

M. Morgan, S. Anders, M. Lawrence, P. Aboyoun, H. Pages, and R. Gentleman. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*, 25(19):2607–2608, Oct 2009. URL `http://www.ncbi.nlm.nih.gov/pubmed/19654119`.

J. L. Mosquera and A. Sanchez-Pla. SerbGO: searching for the best GO tool. *Nucleic Acids Res.*, 36(Web Server issue):W368–371, Jul 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18480123`.

D. M. Muzny, M. N. Bainbridge, K. Chang, H. H. Dinh, and et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, Jul 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22810696`.

A. Nekrutenko and J. Taylor. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat. Rev. Genet.*, 13(9):667–672, Sep 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22898652`.

T. Nikopensius, L. Ambrozaityte, K. U. Ludwig, S. Birnbaum, T. Jagomagi, M. Saag, A. Matuleviciene, L. Linkeviciene, S. Herms, M. Knapp, P. Hoffmann, M. M. Nothen, V. Kucinskas, A. Metspalu, and E. Mangold. Replication of novel susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24 in Estonian and Lithuanian patients. *Am. J. Med. Genet. A*, 149A(11):2551–2553, Nov 2009. URL `http://www.ncbi.nlm.nih.gov/pubmed/19839039`.

D. Nitsch, L. C. Tranchevent, J. P. Goncalves, J. K. Vogt, S. C. Madeira, and Y. Moreau. PINTA: a web server for network-based gene prioritization from expression data. *Nucleic Acids Res.*, 39(Web Server issue):W334–338, Jul 2011.

Oracle. Sun n1 grid engine 6.1 user's guide, 2010. URL `http://docs.oracle.com/cd/E19957-01/820-0699/`.

M. Paganin and A. Ferrando. Molecular pathogenesis and targeted therapies for NOTCH1-induced T-cell acute lymphoblastic leukemia. *Blood Rev.*, 25(2):83–90, Mar 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/20965628`.

C. G. Palii, C. Perez-Iratxeta, Z. Yao, Y. Cao, F. Dai, J. Davison, H. Atkins, D. Allan, F. J. Dilworth, R. Gentleman, S. J. Tapscott, and M. Brand. Differential genomic targeting of the transcription factor TAL1 in alternate haematopoietic lineages. *EMBO J.*, 30(3):494–509, Feb 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21179004`.

C. G. Palii, B. Vulesevic, S. Fraineau, E. Pranckeviciene, A. J. Griffith, A. Chu, H. Faralli, Y. Li, B. McNeill, J. Sun, T. J. Perkins, F. J. Dilworth, C. Perez-Iratxeta, E. J. Suuronen, D. S. Allan, and M. Brand. Trichostatin A enhances vascular repair by injected human endothelial progenitors through increasing the expression of TAL1-dependent genes. *Cell Stem Cell*, 14 (5):644–657, May 2014. URL `http://www.ncbi.nlm.nih.gov/pubmed/24792117`.

R. K. Patel and M. Jain. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE*, 7(2):e30619, 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22312429`.

S. Pepke, B. Wold, and A. Mortazavi. Computation for ChIP-seq and RNA-seq studies. *Nat. Methods*, 6(11 Suppl):22–32, Nov 2009. URL `http://www.ncbi.nlm.nih.gov/pubmed/19844228`.

C. Perez-Iratxeta, P. Bork, and M. A. Andrade. Association of genes to genetically inherited diseases using data mining. *Nat. Genet.*, 31(3):316–319, Jul 2002a. URL `http://www.ncbi.nlm.nih.gov/pubmed/12006977`.

C. Perez-Iratxeta, H. S. Keer, P. Bork, and M. A. Andrade. Computing fuzzy associations for the analysis of biological literature. *BioTechniques*, 32(6):1380–1382, Jun 2002b. URL `http://www.ncbi.nlm.nih.gov/pubmed/12074170`.

C. Perez-Iratxeta, M. Wjst, P. Bork, and M. A. Andrade. G2D: a tool for mining genes associated with disease. *BMC Genet.*, 6:45, 2005. URL `http://www.ncbi.nlm.nih.gov/pubmed/16115313`.

J. W. Pike. Genome-scale techniques highlight the epigenome and redefine fundamental principles of gene regulation. *J. Bone Miner. Res.*, 26(6):1155–1162, Jun 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21611959`.

R. M. Piro and F. Di Cunto. Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS J.*, 279(5):678–696, Mar 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22221742`.

M. Popescu, J. M. Keller, and J. A. Mitchell. Fuzzy measures on the Gene Ontology for gene product similarity. *IEEE/ACM Trans Comput Biol Bioinform*, 3(3):263–274, 2006. URL `http://www.ncbi.nlm.nih.gov/pubmed/17048464`.

E. Portales-Casamar, S. Thongjuea, A. T. Kwon, D. Arenillas, X. Zhao, E. Valen, D. Yusuf, B. Lenhard, W. W. Wasserman, and A. Sandelin. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, 38(Database issue):D105–110, Jan 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/19906716`.

E. Preiksaitiene, J. Kasnauskiene, Z. Ciuladaite, B. Tumiene, P. C. Patsalis, and V. Ku?inskas. Clinical and molecular characterization of a second case of 7p22.1 microduplication. *Am. J. Med. Genet. A*, 158A(5):1200–1203, May 2012.

C. F. Qi, A. Martensson, M. Mattioli, R. Dalla-Favera, V. V. Lobanenkov, and H. C. Morse. CTCF functions as a critical regulator of cell-cycle arrest and death after ligation of the B cell receptor on immature B cells. *Proc. Natl. Acad. Sci. U.S.A.*, 100(2):633–638, Jan 2003. URL `http://www.ncbi.nlm.nih.gov/pubmed/12524457`.

A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, Mar 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/20110278`.

P. Radivojac, K. Peng, W. T. Clark, B. J. Peters, A. Mohan, S. M. Boyle, and S. D. Mooney. An integrated approach to inferring gene-disease associations in humans. *Proteins*, 72(3): 1030–1037, Aug 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18300252`.

P. Ramachandran, G. A. Palidwor, C. J. Porter, and T. J. Perkins. MaSC: mappability-sensitive cross-correlation for estimating mean fragment length of single-end short-read sequencing data. *Bioinformatics*, 29(4):444–450, Feb 2013. URL `http://www.ncbi.nlm.nih.gov/pubmed/23300135`.

D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch, and A. Jimeno. Text processing through Web services: calling Whatizit. *Bioinformatics*, 24(2):296–298, Jan 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18006544`.

P. Resnik. Semantic Similarity in a Taxonomy: an information-based measure and its application to problems of ambiguity in natural language nformation retrieval based on fuzzy associations. *Journal of Artificial Intelligence Research*, 11, 1999.

M. Ridinger-Saison, V. Boeva, P. Rimmele, I. Kulakovskiy, I. Gallais, B. Levavasseur, C. Paccard, P. Legoix-Ne, F. Morle, A. Nicolas, P. Hupe, E. Barillot, F. Moreau-Gachelin, and C. Guillouf. Spi-1/PU.1 activates transcription through clustered DNA occupancy in erythroleukemia. *Nucleic Acids Res.*, 40(18):8927–8941, Oct 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22790984`.

K. R. Rosenbloom, T. R. Dreszer, J. C. Long, V. S. Malladi, C. A. Sloan, B. J. Raney, M. S. Cline, D. Karolchik, G. P. Barber, H. Clawson, M. Diekhans, P. A. Fujita, M. Goldman, R. C. Gravell, R. A. Harte, A. S. Hinrichs, V. M. Kirkup, R. M. Kuhn, K. Learned, M. Maddren, L. R. Meyer, A. Pohl, B. Rhead, M. C. Wong, A. S. Zweig, D. Haussler, and W. J. Kent. ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res.*, 40(Database issue):D912–917, Jan 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22075998`.

D. H. Roukos. Integrated clinical genomics: new horizon for diagnostic and biomarker discoveries in cancer. *Expert Rev. Mol. Diagn.*, 13(1):1–4, Jan 2013. URL `http://www.ncbi.nlm.nih.gov/pubmed/23256697`.

J. Rozowsky, G. Euskirchen, R.K. Auerbach, Z.D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M.B. Gerstein. Peakseq enables systematic scoring of chip-seq experiments relative to controls. *Nature biotechnology*, 7(1):66–75, Jan 2009. URL `http://www.ncbi.nlm.nih.gov/pubmed/19122651`.

M. Ruffalo, T. LaFramboise, and M. Koyuturk. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, 27(20):2790–2796, Oct 2011. URL http://www.ncbi.nlm.nih.gov/pubmed/21856737.

E. Sayers. The e-utilities in-depth: Parameters, syntax and more, 2009. URL http://www.ncbi.nlm.nih.gov/books/NBK25499/.

M. A. Schaub, A. P. Boyle, A. Kundaje, S. Batzoglou, and M. Snyder. Linking disease associations with regulatory information in the human genome. *Genome Res.*, 22(9):1748–1759, Sep 2012. URL http://www.ncbi.nlm.nih.gov/pubmed/22955986.

R. Schmieder and R. Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864, Mar 2011. URL http://www.ncbi.nlm.nih.gov/pubmed/21278185.

S. Sebastian, H. Faralli, Z. Yao, P. Rakopoulos, C. Palii, Y. Cao, K. Singh, Q. C. Liu, A. Chu, A. Aziz, M. Brand, S. J. Tapscott, and F. J. Dilworth. Tissue-specific splicing of a ubiquitously expressed transcription factor is essential for muscle differentiation. *Genes Dev.*, 27(11):1247–1259, Jun 2013. URL http://www.ncbi.nlm.nih.gov/pubmed/23723416.

H. Shatkay and M. Craven. *Mining the Biomedical Literature*. MIT Press, 2012.

J. Shendure and E. Lieberman Aiden. The expanding scope of DNA sequencing. *Nat. Biotechnol.*, 30(11):1084–1094, Nov 2012. URL http://www.ncbi.nlm.nih.gov/pubmed/23138308.

N. R. Smalheiser and D. R. Swanson. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput Methods Programs Biomed*, 57(3):149–153, Nov 1998. URL http://www.ncbi.nlm.nih.gov/pubmed/9822851.

N. R. Smalheiser, V. I. Torvik, and W. Zhou. Arrowsmith two-node search interface: a tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. *Comput*

Methods Programs Biomed, 94(2):190–197, May 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19185946.

B. Smith, M. Ashburner, C. Rosse, J. Bard, and et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, 25(11):1251–1255, Nov 2007. URL http://www.ncbi.nlm.nih.gov/pubmed/17989687.

C. L. Smith and J. T. Eppig. The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mamm. Genome*, 23(9-10):653–668, Oct 2012. URL http://www.ncbi.nlm.nih.gov/pubmed/22961259.

K. Struhl. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*, 98(1):1–4, Jul 1999. URL http://www.ncbi.nlm.nih.gov/pubmed/10412974.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, 102(43):15545–15550, Oct 2005. URL http://www.ncbi.nlm.nih.gov/pubmed/16199517.

S. Suzuki, M. L. Marazita, M. E. Cooper, N. Miwa, A. Hing, A. Jugessur, N. Natsume, K. Shimozato, N. Ohbayashi, Y. Suzuki, T. Niimi, K. Minami, M. Yamamoto, T. J. Altannamar, T. Erkhembaatar, H. Furukawa, S. Daack-Hirsch, J. L'heureux, C. A. Brandon, S. M. Weinberg, K. Neiswanger, F. W. Deleyiannis, J. E. de Salamanca, A. R. Vieira, A. C. Lidral, J. F. Martin, and J. C. Murray. Mutations in BMP4 are associated with subepithelial, microform, and overt cleft lip. *Am. J. Hum. Genet.*, 84(3):406–411, Mar 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19249007.

D. R. Swanson. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.*, 30(1):7–18, 1986. URL http://www.ncbi.nlm.nih.gov/pubmed/3797213.

A. M. Szalkowski and C. D. Schmid. Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts. *Brief. Bioinformatics*, 12(6):626–633, Nov 2011.

M. Szekeres, G. L. Nadasy, G. Turu, E. Soltesz-Katona, Z. E. Toth, A. Balla, K. J. Catt, and L. Hunyady. Angiotensin II induces vascular endocannabinoid release, which attenuates its vasoconstrictor effect via CB1 cannabinoid receptors. *J. Biol. Chem.*, 287(37):31540–31550, Sep 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22787147`.

D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. von Mering. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, 39(Database issue):D561–568, Jan 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21045058`.

B. Tang, P. Di Lena, L. Schaffer, S. R. Head, P. Baldi, and E. A. Thomas. Genome-wide identification of Bcl11b gene targets reveals role in brain-derived neurotrophic factor signaling. *PLoS ONE*, 6(9):e23691, 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21912641`.

A. K. Tewari, G. G. Yardimci, Y. Shibata, N. C. Sheffield, L. Song, B. S. Taylor, S. G. Georgiev, G. A. Coetzee, U. Ohler, T. S. Furey, G. E. Crawford, and P. G. Febbo. Chromatin accessibility reveals insights into androgen receptor activation and transcriptional specificity. *Genome Biol.*, 13(10):R88, Oct 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/23034120`.

P. Thompson, R. Nawaz, J. McNaught, and S. Ananiadou. Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, 12:393, 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21985429`.

N. Tiffin, M. A. Andrade-Navarro, and C. Perez-Iratxeta. Linking genes to diseases: it's all in the data. *Genome Med*, 1(8):77, 2009. URL `http://www.ncbi.nlm.nih.gov/pubmed/19678910`.

L. Tranchevent. *Gene prioritization through genomic data fusion.* PhD thesis, Faculty of Engineering, K.U.Leuven (Leuven, Belgium), 2011.

L. C. Tranchevent, R. Barriot, S. Yu, S. Van Vooren, P. Van Loo, B. Coessens, B. De Moor, S. Aerts, and Y. Moreau. ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res.*, 36(Web Server issue):W377–384, Jul 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18508807`.

L. C. Tranchevent, F. B. Capdevila, D. Nitsch, B. De Moor, P. De Causmaecker, and Y. Moreau. A guide to web tools to prioritize candidate genes. *Brief. Bioinformatics*, 12(1):22–32, Jan 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21278374`.

N. Turenne, E. Tiys, V. Ivanisenko, N. Yudin, E. Ignatieva, D. Valour, S. A. Degrelle, and I. Hue. Finding biomarkers in non-model species: literature mining of transcription factors involved in bovine embryo development. *BioData Min*, 5(1):12, 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22931563`.

UCSC-Genome-Bioinformatics. Human feb. 2009 (grch37/hg19) (hg19), 2009. URL `http://genome.ucsc.edu/goldenPath/releaseLog.html`.

G. Valentini, A. Paccanaro, H. Caniza, A. E. Romero, and M. Re. An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods. *Artif Intell Med*, 61(2):63–78, Jun 2014.

M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. Leunissen. A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.*, 14(5):535–542, May 2006. URL `http://www.ncbi.nlm.nih.gov/pubmed/16493445`.

P. Van Loo and P. Marynen. Computational methods for the detection of cis-regulatory modules. *Brief. Bioinformatics*, 10(5):509–524, Sep 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19498042.

P. Van Vlierberghe and A. Ferrando. The molecular basis of T cell acute lymphoblastic leukemia. *J. Clin. Invest.*, 122(10):3398–3406, Oct 2012. URL http://www.ncbi.nlm.nih.gov/pubmed/23023710.

Y. Wakabayashi, H. Watanabe, J. Inoue, N. Takeda, J. Sakata, Y. Mishima, J. Hitomi, T. Yamamoto, M. Utsuyama, O. Niwa, S. Aizawa, and R. Kominami. Bcl11b is required for differentiation and survival of alphabeta T lymphocytes. *Nat. Immunol.*, 4(6):533–539, Jun 2003. URL http://www.ncbi.nlm.nih.gov/pubmed/12717433.

P. I. Wang and E. M. Marcotte. It's the machine that matters: Predicting gene function and phenotype from protein networks. *J Proteomics*, 73(11):2277–2289, Oct 2010. URL http://www.ncbi.nlm.nih.gov/pubmed/20637909.

W. W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, 5(4):276–287, Apr 2004. URL http://www.ncbi.nlm.nih.gov/pubmed/15131651.

A. J. Webb, G. A. Thorisson, and A. J. Brookes. An informatics project and online "Knowledge Centre" supporting modern genotype-to-phenotype research. *Hum. Mutat.*, 32(5):543–550, May 2011. URL http://www.ncbi.nlm.nih.gov/pubmed/21438073.

H. Wickham. stringr: Make it easier to work with strings, 2012. URL http://cran.r-project.org/web/packages/stringr/index.html.

E. G. Wilbanks and M. T. Facciotti. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE*, 5(7):e11471, 2010. URL http://www.ncbi.nlm.nih.gov/pubmed/20628599.

E. Wingender, P. Dietze, H. Karas, and R. Knuppel. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, 24(1):238–241, Jan 1996. URL http://www.ncbi.nlm.nih.gov/pubmed/8594589.

K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, S. Soiland-Reyes, I. Dunlop, A. Nenadic, P. Fisher, J. Bhagat, K. Belhajjame, F. Bacall, A. Hardisty, A. Nieva de la Hidalga, M. P. Balcazar Vargas, S. Sufi, and C. Goble. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res.*, 41(Web Server issue):W557–561, Jul 2013. URL http://www.ncbi.nlm.nih.gov/pubmed/23640334.

Z. Yao, A. P. Fong, Y. Cao, W. L. Ruzzo, R. C. Gentleman, and S. J. Tapscott. Comparison of endogenous and overexpressed MyoD shows enhanced binding of physiologically bound sites. *Skelet Muscle*, 3(1):8, 2013. URL http://www.ncbi.nlm.nih.gov/pubmed/23566431.

Y. Yoshida, Y. Makita, N. Heida, S. Asano, A. Matsushima, M. Ishii, Y. Mochizuki, H. Masuya, S. Wakana, N. Kobayashi, and T. Toyoda. PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning. *Nucleic Acids Res.*, 37(Web Server issue):W147–152, Jul 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19468046.

S. Yu, S. Van Vooren, L. C. Tranchevent, B. De Moor, and Y. Moreau. Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining. *Bioinformatics*, 24(16):i119–125, Aug 2008a. URL http://www.ncbi.nlm.nih.gov/pubmed/18689812.

W. Yu, A. Wulf, T. Liu, M. J. Khoury, and M. Gwinn. Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases. *BMC Bioinformatics*, 9:528, 2008b. URL `http://www.ncbi.nlm.nih.gov/pubmed/19063745`.

L. A. Zadeh. From computing with numbers to computing with words. From manipulation of measurements to manipulation of perceptions. *Ann. N. Y. Acad. Sci.*, 929:221–252, Apr 2001. URL `http://www.ncbi.nlm.nih.gov/pubmed/11357866`.

B. R. Zeeberg, W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett, and J. N. Weinstein. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, 4(4):R28, 2003. URL `http://www.ncbi.nlm.nih.gov/pubmed/12702209`.

J. A. Zhang, A. Mortazavi, B. A. Williams, B. J. Wold, and E. V. Rothenberg. Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity. *Cell*, 149(2):467–482, Apr 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22500808`.

Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9(9):R137, 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18798982`.

X. Zhou, S. Bao, B. Wang, X. Zhang, and Y. Song. Short read mapping for exome sequencing. In *Deep Sequencing Data Analysis, Methods in Molecular Biology*, pages 93–111. Springer, 2013.

L. J. Zhu, C. Gazin, N. D. Lawson, H. Pages, S. M. Lin, D. S. Lapointe, and M. R. Green. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*, 11:237, 2010a. URL `http://www.ncbi.nlm.nih.gov/pubmed/20459804`.

Q. Zhu, M. S. Lajiness, Y. Ding, and D. J. Wild. WENDI: A tool for finding non-obvious relationships between compounds and biological properties, genes, diseases and scholarly publications. *J Cheminform*, 2:6, 2010b. URL `http://www.ncbi.nlm.nih.gov/pubmed/20727184`.

H.-J. Zimmermann. Fuzzy set theory. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3):317–332, 2010. ISSN 1939-0068. doi: 10.1002/wics.82. URL `http://dx.doi.org/10.1002/wics.82`.

H.J. Zimmermann. *Fuzzy Set Theory and its applications*. Kluwer Academic Publishers, 1996.

Appendix A. Supplementary Material for Chapter 3

This Appendix contains supplementary material for Chapter 3 regarding technical and user data.

Backend Database of Gene Prioritization Algorithms. The developed prioritization algorithm uses a local database to aid it in computation of the links between genes and phenotypes. It stores information about co-occurrences of MeSH annotations and GO annotations in MEDLINE and Gene Ontology, as well as information about ancestors and descendants for GO anotations. The data sources used to create the database are listed in Section A.1. The database content is shown in Section A.2.

The gene prioritization has two parameters which set the minimum thresholds on the strengths of the connections between phenotypes defined by MeSH terms and chemicals defined by MeSH D terms; and between chemicals (MeSH D terms) and GO annotations. The meaning of the parameters and principles of their selection are explained in Section A.3.

A comparison between the proposed algorithm and CANDID tool was performed by prioritizing human disease genes. The comparison of the ranking by both algorithms is presented in Section A.4.

Performance of several benchmark cases were discussed throughout the thesis. The details about these benchmark cases are presented here. The "**Respiration**" phenotype is outlined in

Section A.5. The "**DNA Repair**" phenotype is presented in Section A.6. The "**Autophagy**" phenotype is discussed in Section A.7. The "**Vasoconstriction**" phenotype is presented in Section A.8.

The created prioritization algorithm can is available as a web application. The user interface and manual are presented in Section A.9.

A.1 Data sources Used to Create a Local Database for the Algorithm

The data sources used to create the backend database are listed in Table A.1.

TABLE A.1: Data Sources Used to Create the Database for the Gene Prioritization Algorithm

Database	**Data Source**	**Use in Prioritization Algorithm**
MeSH (2012 Edition)	mtree2012.bin	To have the full list of the MeSH terms with the corresponding category identifier. In our implementation we used the A,B,C,D and G categories. These MeSH terms were used to retrieve the corresponding PubMed identifiers of the articles having co-occuring MeSH terms.
PubMed (as of March 2013)	Articles and annotations retrieved by *E-utilities*	The article annotations were used to create the table of the PMID counts for each pair of the MeSH annotations that co-occurred in the article annotations.
NCBI Entrez Gene	gene2go	To collect the annotated genes of the human, mouse and fly together with their GO annotations.
	gene2pubmed	To retrieve the GO annotations co-occurring with the MeSH D terms in the articles associated with the genes.
	gene.info	To create the gene information table for the annotated genes of human, mouse and fly.
	homolo.gene	To create the records containing the homologous genes of the three organisms and their GO annotations.
Gene Ontology (as of March 2013)	geneontology.obo flat file	To create the table containing pairs of the GO annotations and their common ancestor for the three organisms and to compute the information content of each GO annotation.

A.2 Contents of the Local Database of the Prioritization Algorithm

The content of the created database is summarized in Table A.2.

TABLE A.2: Tables in the Database

Content of the Table	**Size of the Table**
The table of frequencies of PubMed articles annotated by the pair of the MeSH terms. The pairs of the MeSH terms consist of the terms in the A, C, D and G categories coupled with all terms from the D category. Example of such pair is "**Cell Fusion**" and "**Viral Fusion Proteins**". In order to create this table the E-utils were used to scan the whole MEDLINE.	9,725,157 pairs .
The table of the associations between the functional gene annotations and the MeSH D terms. Example of such pair is "**Viral Fusion Proteins**" and "**protein transport within lipid bilayer** ". Each pair has the corresponding genes in which those annotations occur together. The table was created by using NCBI Entrez Gene [Maglott et al., 2011] databases. The *gene2go* table was used to retrieve all annotated human, mouse and fly genes and their GO annotations. The *gene2pubmed* table was used to find the genes having the particular GO annotation and associated with the articles, annotated by the particular MeSH D term.	14,225,540 pairs.
The table containing gene information from the *gene.info* for human, mouse and fly genes.	124,785 genes.
The table storing the mapping between the homologous genes of the three organisms. The *homolo.gene* was used to create this table.	742,162 mappings.
The table containing all pairs of the GO annotations with the closest common ancestor of those pairs of the human, mouse and fly genes. The GO annotations and their descriptions for the three organisms were retrieved from the original Gene Ontology database.	43,873,781 pairs.
The table with the computed information content for each unique GO annotation from the table of the ancestors.	15,160 records.
The table mapping the gene symbols and their synonyms to the Entrez Gene Identifiers.	373,645 mappings.
Index, helper and temporary tables.	Variable.

A.3 Parameters of Literature-Based Prioritization Algorithm

The literature-based prioritization algorithm depends on two parameters D threshold and GO threshold. These parameters determine the minimum fuzzy membership values of the fuzzy sets defined on MeSH terms and GO annotations. These parameters determine the α-cuts of the fuzzy sets (explained in Section 3.2.3.4). Values of these parameters control the formation of the links between the phenotype and the GO annotation. A value of the D threshold (t_{GD} in Equation 3.12) enables a filter which discards the links between the MeSH G and MeSH D terms with fuzzy membership below that threshold value. High values of the D threshold make sure that only the strongest connections between the phenotype and the chemicals are used in the inference. The low values allow weak connections among which the spurious noninformative connections appear. Similarly a value of the GO threshold (t_{DGO} in Equation 3.12) enables a filter which discards the links between the MeSH D terms and GO annotations that have fuzzy membership below that threshold. The optimal values of the D and the GO thresholds were determined using the benchmark.

As a global measure of performance for fixed parameter values a percentage of the benchmark cases with significant predictions (empirical p value ≤ 0.05) was used. Computations were performed on a small grid of the parameter values. The plots of this global performance measure over the grid of the parameters is presented in Figure A.3. The Y-axis represents a percentage of the benchmark sets in which the known gene set had significant predictions indicated by the empirical p value ≤ 0.05 in the Top1% of the genome (as explained in previous Section **??**). As D and GO thresholds increase, less information is included in the ranking scheme and we observe a diminishing performance of the algorithm. The panel on the right represents an enlarged section of the panel on the left (GO threshold between 0.001 and 0.1). The best performance is achieved if the D threshold equals 0.0001 and the GO threshold is less than 0.001. However, with such parameter values the real time response of the web application of the user interface is very slow.

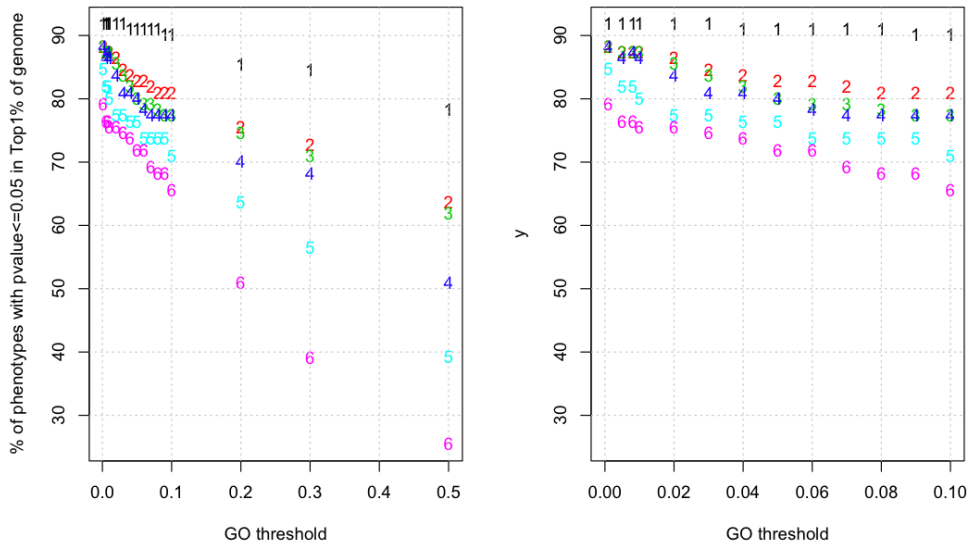


FIGURE A.3: Global Performance of Gene Ranking Algorithm in the Benchmark with Respect to Different Values of the D Threshold and GO Threshold. X Axis Represents Levels of the GO Threshold. The Levels of the D Threshold (DT) Are Symbol Coded: **1** DT=0.001, **2** DT=0.0008, **3** DT=0.001, **4** DT=0.002, **5** DT=0.005 and **6** DT=0.01. Y Axis Represents the Percentage of the Benchmark Sets Having Significant Predictions

To ensure a balance between a performance and speed the following thresholds were chosen: D threshold = 0.001 and GO threshold = 0.01. With these values there were significant predictions in 87.3% of the benchmark cases within the Top 1% of the ranked genome. The results of the benchmark will be discussed with respect to the fixed D and GO thresholds 0.001 and 0.01 respectively.

A.4 Comparison of Literature-Based Prioritization Algorithm with CANDID Tool

The CANDID tool [Hutz et al., 2008] is closest in functionality to the literature-based prioritization algorithm developed in this thesis. CANDID is designed to work with human genes. In order to understand how these algorithms compare to each other several gene lists from the CANDID publication [Hutz et al., 2008] associated with human diseases were explored. The genes, that were associated with the disease which has a corresponding MeSH term were selected. All these single genes from the CANDID publication are already known to be associated with the diseases.

Genes of in the whole human genome were ranked by CANDID and the developed algorithm with respect to the given disease phenotypes. The positions of the selected genes in the ranked genome lists were identified. The ranking results of CANDID and the algorithm proposed in this thesis are shown in Table A.4. In the table the genes that had higher rank out of the two are presented in bold.

The 9 genes out of the 23 were ranked higher by CANDID. The developed algorithm ranked the 14 genes higher than CANDID. These ranking results serve exploratory purpose. No superiority of one method versus the other is claimed based on these ranking results because the algorithms use different data sources and different computational approaches. Comparison of the algorithms should be based on a sufficiently large and carefully designed benchmark test set, which at the current time is not available for the community yet [Moreau and Tranchevent, 2012].

TABLE A.4: Ranking of the Human Disease Genes by the Proposed Algorithm and CANDID Tool. Numbers in Bold Represent a Higher Rank out of Two: Either CANDID or the Developed Algorithm

Disease	**Gene**	**Gene Description**	**CANDID Rank**	**Literature Based Algorithm Rank**
Colon cancer	SMAD7	SMAD family member 7	954	**458**
Breast cancer	BRIP1	BRCA1 interacting protein C-terminal helicase 1	**1912**	5292
	CASP8	caspase 8 apoptosis-related cysteine peptidase	**67**	1382
	FGFR2	fibroblast growth factor receptor 2	5404	**77**
	LSP1	lymphocyte-specific protein 1	**7108**	10047
	MAP3K1	mitogen-activated protein kinase kinase kinase 1 E3 ubiquitin protein ligase	**359**	575
	PALB2	partner and localizer of BRCA2	11251	**5085**
	TOX3	TOX high mobility group box family member	10741	**166**
Alzheimer's disease	PLAU	plasminogen activator urokinase	115	**90**
	SORL1	sortilin-related receptor L(DLR class) A repeats containing	**148**	1330
	OPCML	opioid binding protein/cell adhesion molecule-like	**486**	718
Crohn's disease	ATG16L1	autophagy related 16-like 1 (S. cerevisiae)	5657	**3152**
	FAM92B	family with sequence similarity 92 member B	14848	**10978**
	NCF4	neutrophil cytosolic factor 4 40kDa	3566	**2744**
	PHOX2B	paired-like homeobox 2b	3349	**22**
Type 2 Diabetes	CDKAL1	CDK5 regulatory subunit associated protein 1-like 1	**6107**	12335
	CDKN2A	cyclin-dependent kinase inhibitor 2A	2755	**1528**
	CDKN2B	cyclin-dependent kinase inhibitor 2B (p15 inhibits CDK4)	16521	**1203**
	FTO	fat mass and obesity associated	7615	**1834**
	HHEX	hematopoietically expressed homeobox	4338	**152**
	IDE	insulin-degrading enzyme	**129**	809
	SLC30A8	solute carrier family 30 (zinc transporter) member 8	**2829**	3080
Rheumatoid arthritis	C5	complement component 5	231	**45**
	TRAF1	TNF receptor-associated factor 1	**389**	1048

A.5 Details of the "Respiration" Phenotype

The "**Respiration**" phenotype in MeSH is defined as *"the act of breathing with the lungs, consisting of inhalation, or the taking into the lungs of the ambient air, and of exhalation, or the expelling of the modified air which contains more carbon dioxide than the air taken in"*. The test set related to this phenotype consists of 22 genes from MGI database. The three top ranking genes and the strongest relationships between the chemicals and GO annotations are summarized in Table A.5.

TABLE A.5: The Benchmark Details for the "**Respiration**" Phenotype

Gene	**Description**	**Gene Score (Rank)**	**GO Annotation**	**GO Score**	**MeSH D**
Grin1	Glutamate receptor ionotropic NMDA1 (zeta 1)	0.4873 (120)	Pons maturation	0.0449	Anesthetics, Dissociative
Mecp2	Methyl CpG binding protein 2	1.2669 (312)	neurological system process involved in regulation of systemic arterial blood pressure	0.0347	Oxygen
Ache	Acetylcholinesterase	1.4293 (352)	negative regulation of synaptic transmission, cholinergic	0.0624	Oxygen

Relation GO - Phenotype	**GO Annotation**	**MeSH D**	**Relation MeSH D - Phenotype**
0.177	bradykinin biosyntetic process	Xylazine	0.177
0.177	tissue kallikrein-kinin cascade	Xylazine	0.177
0.1492	lobar bronchus epitelium develpment	Buthorphanol	0.1492
0.1397	negative regulation of uterine smooth muscle contractions	Medetomidine	0.1856
0.1073	regulation of leukocyte mediated cytotoxicity	Helium	0.1073
0.0965	albendazole monooxygenase activity	Oxymorphone	0.0965
0.0965	(R)-limonene 6-monooxygenase activity	Oxymorphone	0.0965
0.0965	guinine 3-monooxygenase activity	Oxymorphone	0.0965

The MeSH D terms which have strongest relationships with the "**Respiration**" phenotype are the "**Medetomidine**", "**Xylazine**", "**Butorphanol**", "**Nalbuphine**", "**Helium**", "**Oxymorphone**", "**Anesthetics**" and "**Dissociative**". The connection between "**Oxygen**" and "**Respiration**" is rather weak compared to the other chemicals linked to "**Respiration**".

A.6 Details of the "DNA Repair" Phenotype

The "**DNA Repair**" phenotype and genes associated with it are very well characterized. "**DNA Repair**" was the best performing phenotype in the benchmark.

TABLE A.6: The Benchmark Details for the "**DNA Repair**" Phenotype

Gene	**Description**	**Gene Score (Rank)**	**GO Annotation**	**GO Score**	**MeSH D**
Ogg1	8-oxoguanine DNA-glycosylase 1	0.0203 (5)	depurination	0.5202	DNA Glycosylases
Sirt1	sirtuin 1 (silent mating type information regulation 2, homolog) 1 (S cerevisiae)	0.0528 (13)	negative regulation of peptidyl-lysine acetylation	0.4249	DNA-Activated Protein Kinase
Lig4	ligase IV, DNA, ATP-dependent	0.0809 (29)	DNA ligation involved in DNA recombination	0.4249	DNA-Activated Protein Kinase

The top gene annotations linked to "**DNA Repair**" refer to biochemical reactions taking place during DNA modifications: "**depurination, negative regulation of peptidyl-lysine acetylation**", "**DNA ligation involved in DNA recombination**". These GO annotations are strongly related to "**DNA Glycosylases**" in 1840 MEDLNE articles annotated by both the "**DNA Repair**" and the "**DNA Glycosylases**". The "**DNA Glycosylases**" represent a family of DNA repair enzymes that recognize damaged nucleotide bases and remove them by base excision repair. The mentioned GO annotations are also strongly related to "**DNA-Activated Protein Kinase**" in 574 MEDLINE articles. "**DNA-Activated Protein Kinase**" is involved

in phosphorylation of DNA binding protein substrates. Other top MeSH D terms strongly connected to "**DNA repair**" are: "**Xeroderma Pigmentosum Group D Protein**" linked to "**DNA repair**" by the 370 articles, "**Deoxyribonuclease IV Phage T4-Induced**" linked by 318 articles, "**Thymine DNA Glycosylase**" linked by 55 articles, "**MutS DNA Mismatch-Binding Protein**" related to the "**DNA repair**" by 291 articles and "**Rad52 DNA Repair and Recombination Protein**" related by 255 articles. All of these molecular entities are directly involved in DNA modification processes such as cleavage and excision. The "**DNA Repair**" phenotype has 40730 associated articles in total (as of September 2013). It is extensively studied in literature and linked to very specific bio-molecules.

A.7 Details of the "Autophagy" Phenotype

The gene list associated with "**Autophagy**" has 18 genes. Neither of those genes is within the Top 1% of the ranked genome. In the Mammalian Phenotype Ontology the "**abnormal autophagy**" phenotype is defined as: *"Abnormal catabolic process involving the degradation of a cell's own components through the lysosomal machinery".* The "**Autophagy**" phenotype in MeSH Vocabulary is defined as: *"The segregation and degradation of damaged or unwanted cytoplasmic constituents by autophagic vacuoles (cytolysosomes) composed of lysosomes containing cellular components in the process of digestion; it plays an important role in the biological metamorphosis of amphibians, in the removal of bone by osteoclasts and in the degradation of normal cell components in nutritional deficiency states".*

The genes associated to the autophagy phenotype in Mouse Genome Informatics database: Atg16l1, Atg4b, Atg4c, Atg9a, Becn1, Bnip3l, Cisd2, Clcn7, Grid2, Ido1, Lrrk2, Mtap1s, Nhlrc1, Nod2, Oxct1, Pmp22, Pthlh, Ulk1 and Vcp had low ranks. Several of those genes are listed in

Table A.7. The strongest relationships, the MeSH D terms linked to the "**Autophagy**" in MEDLINE comprise "**Class III Phosphatidylinositol 3-Kinases**" (46 articles), "**Lysosomal-Associated Membrane Protein 2**" (51 articles) and "**Ubiquitinated Proteins**" (26 articles). These MeSH D terms designate very broad class of processes. They are strongly linked to the GO annotations representing many different functions including the functions directly related to the autophagy phenotype.

For example, the "**Class III Phosphatidylinositol 3-Kinases**" term has the strongest relationship with "**diaphragm morphogenesis, regulation of G1/S transition checkpoint** " and "**regulation of cytokine secretion involved in the immune response** ".

The "**Ubiquitinated Proteins**" term has strongest relationships with the GO annotations such as "**M Phase, regulation of histone H3-K36 methylation** ", "**regulation of transcription from RNA polymerase II promoter by glucose**", "**condensin core heterodimer**", "**branched-chain amino acid catabolic process to alcohol via Ehrlich pathway**" and 443 more annotations that describe very heterogeneous cellular processes.

The phenotypes linked to ubiquitous bio-molecules such as "**Class III Phosphatidylinositol 3-Kinases**" and "**Ubiquitinated Proteins**" will favor the multifunctional genes in prioritization. For example the Gata4 gene appears top first in the ranked whole genome with respect to the "**Autophagy**" phenotype. The Gata4 gene encodes a transcriptional activator. It regulates genes involved in embryogenesis and in myocardial differentiation. The strongest GO annotation for Gata4 which placed that gene at the top is "**diaphragm morphogenesis** " most strongly linked to "**Class III Phosphatidylinositol 3-Kinases**". The Knnc4 gene having 32 annotations appears in the top second position of the ranked whole genome. The highest scoring annotation for this gene is "**intermediate conductance calcium-activated potassium channel activity** " also having the strongest relation to "**Class III Phosphatidylinositol 3-Kinases**".

TABLE A.7: The Prioritization Result for the "**Autophagy**" Benchmark Phenotype

Gene	**Description**	**Gene Score (Rank)**	**GO Annotation**	**GO Score**	**MeSH D**
Atg9a	Autophagy related 9A	12.75 (4857)	Protein localization to Golgi aparatus	0.0337	Ubiquitinated Proteins
Atg4b	Autophagy related 4B cysteine peptidase	13.91 (5365)	Positive regulation of protein catabolic process	0.0326	Ubiquitinated Proteins
Atg16l1	Autophagy related 16-like 1 (S.cerevisiae)	19.45 (4791)	Protein binding	0.0286	Ubiquitinated Proteins
Atg4c	Autophagy related 4C cysteine peptidase	26.22 (8405)	Autophagic vacuole assembly	0.0286	Class III Phosphatidylinositol 3-Kinases

A.8 Details of the "Vasoconstriction" Phenotype

The "**Vasoconstriction**" in the MeSH vocabulary is defined as: *"The physiological narrowing of blood vessels by contraction of the vascular smooth muscle"*.

The gene list associated with the "**Vasoconstriction**" phenotype has 11 genes out of which the 6 genes (66%) are placed within the Top 1% of the ranked genome by our algorithm. Other genes appear within the Top 5% of the ranked genome except of Kcnmb1 potassium large conductance calcium-activated channel subfamily M beta member 1. The top ranking genes are listed in Table A.8. The highest scoring GO annotations have strongest links to "**Vasoconstriction Agents**" drugs that are used to cause constriction of the blood vessels. The other bio-molecules most strongly connected to "**Vasoconstriction**" are " **Adrenergic alpha-1 Receptor Agonists, Receptor**", "**Endothelin B, Receptors**", "**Purinergic P2X1**" and "**Endothelins**". The top scoring GO annotations the "**vasoconstriction of artery involved in baroreceptor response to lowering of systemic arterial blood pressure**", "**regulation of systemic arterial blood pressure by norepinephrine-epinephrine**". The "**Vasoconstriction Agents**" term contributed most to the weights of the the top ranking whole genome genes not included in the

TABLE A.8: The Strongest Gene Candidates Linked to the "**Vasoconstriction**" Phenotype

Gene	**Description**	**Gene Score (Rank)**	**GO Annotation**	**GO Score**	**MeSH D**
Agtr2	Angiotensin II receptor type 2	0.0528 (17)	Aldosterone secretion	0.5123	Vasoconstrictor Agents
Cav1	Caveolin 1 caveolae protein	0.199 (63)	regulation of the force of heart contraction by chemical signalling	0.1142	Vasoconstrictor Agents
Uts2r	Urotensin 2 receptor	0.0203 (64)	Urotensin II receptor activity	0.1142	Vasoconstrictor Agents

benchmark gene list such as Agt (angiotensinogen (serpin peptidase inhibitor clade A member 8)) and Ace (angiotensin I converting enzyme (peptidyl-dipeptidase A) 1).

A.9 User Manual

The algorithms have been made available through the links at `http://figshare.com/PhenotypeLinks`. To use the developed algorithms a user provides an organism (human, mouse or fly), a keyword describing the phenotype of interest and/or set of genes already known to be related to the phenotype. The whole genomes of human *Homo sapiens*, mouse *Mus musculus* or fly *D.melanogaster* are ranked according to the strength of each gene association with the phenotype of interest. Optionally, a list of candidate genes to be prioritized with respect to the given phenotype can be provided. These data sources can be combined: literature and homology between organisms. The phenotype definition may be augmented by the training genes if available, which represents an additional data source in ranking. The whole genome of the selected organism is ranked separately for each selected data source. At the output the web tool displays the genome genes and candidates among them sorted by the combined gene score. It is explained in detail in the thesis Section 3.1.4. Next to the score the position of the gene in the whole ranked genome is shown. The user can browse all established associations and examine which concepts have been linked and the strength of the connections between them. The user interface is summarized in the following manual.

User Manual

Table of Contents

1. Introduction to the Phenotype Links Tool

Back to the Table of Contents

Often molecular biologists have to analyze poorly characterized gene lists obtained in Microarray or Next Generation Sequencing or other experiments. The smaller number of the candidate genes, strongly related to the phenotype of interest, should be selected from the large list. Most of the current gene prioritization tools for phenotype definition use genes with known functions for phenotype definition. Our Phenotype Links Tool provides the functionality absent in the currently available gene prioritization tools such as the phenotype definition by the meaningful keyword. To define the phenotypes, we use the Medical Subject Header (MeSH) terms. The MeSH terms are created by the National Library of Medicine (NLM) to annotate a content of the articles in the MEDLINE database by manual curation. Our system deduces the relationships between the phenotype and the functional Gene Ontology annotations of the genes. The strength of these relationships is used to prioritize the genes in the list of candidates. We utilized a computational algorithm similar to the algorithm in the G2D (genes to diseases) system.

Our algorithm assigns weights to the Gene ontology (GO) annotations that express a strength of the connection between a gene feature and the phenotype. Large weights encode strong connections. Annotated genes are ranked by the weights of their annotations from the most promising to the least promising. Towards the top of the ranked gene list we expect to find genes that are strongly connected to the phenotype, based on the current state of knowledge in the literature. The Homology between the genes of the different organisms can be used as an additional data source. If so, then the weighted gene annotations from the other organisms are included into the prioritization process. The user can also define the phenotype by a the context genes with the known functions. The annotations of the context genes are used to weight all other GO annotations by the semantic Resnik similarity.

We rank the whole genome of the selected organism separately for each selected data source. The gene ranks are transformed into a relative score also known as a rank ratio. The lesser is the relative score, the better the candidate gene is. A combined relative score of the gene is the best relative score out of all scores corresponding to the provided data sources. As a result of the prioritization, the genes sorted by the combined relative score are listed.

We tested our tool by using the gene sets from Mouse Genome Informatics (MGI) database. For the algorithmic details, estimation of the performance and a list of references we refer the user to the Technical Manual. The User Manual describes the main features of the Phenotype Links Tool user interface.

2. User Input

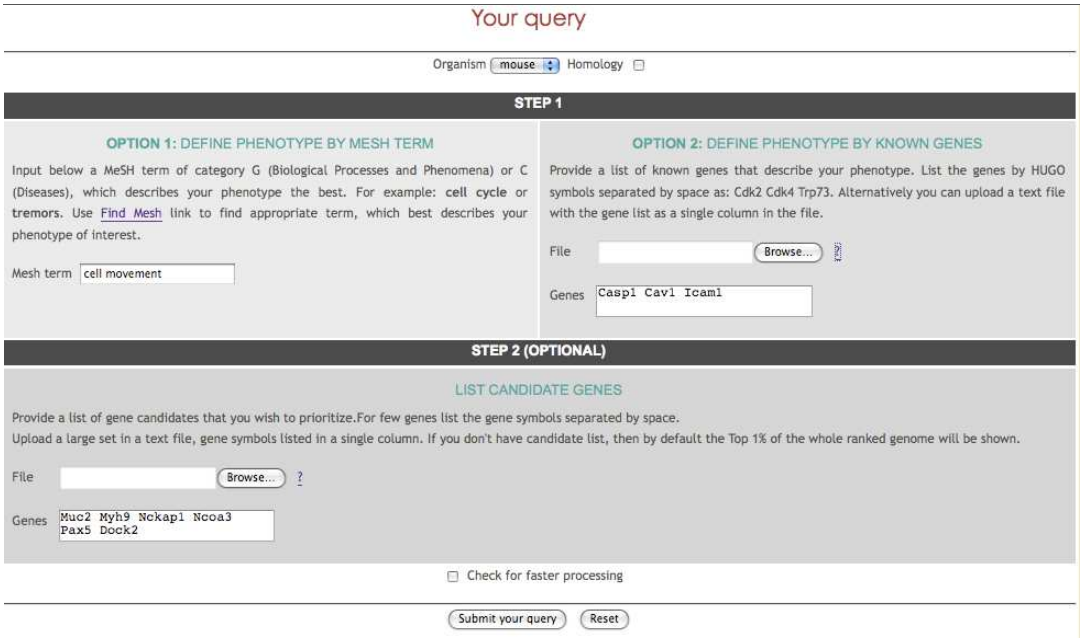
Back to the Table of Contents

The user input interface is shown in Figure 1. To explore the tool you can click on the link Cell Movement phenotype example, in which the input fields will be filled automatically, then submit the example and explore the output. Our tool requires user to provide the phenotype definition either by the MeSH term or by the known genes or both. The mandatory elements of the input consist of:

1. the organism,
2. whether the homology information is included,
3. the phenotype description.

The phenotype description by the MeSH term consists of the text phrase, whch matches the MeSH term exactly. Our tool supports the phenotypes described by the MeSH terms of the categories G (Biological Phenomena and Processes), C (Diseases), D (Chemicals and Drugs) and A (Anatomy) in 2012 edition of the MeSH Vocabulary. The alphabetically ordered summary of the MeSH keywords that can be used for the phenotype definition can be reviewed here.

Figure 1. User Input Area



Select the organism from the list field. If you choose to use homology information btween the organisms, check the Homology box.

In the Step 1 field OPTION 1 type the text phrase corresponding to the MeSH term, which describes your phenotype.

In the OPTION 2 list the context genes separated by space, defining your phenotype. If the list of the context genes is large, then you can upload the text file with the gene symbols. The text file has only one column as shown in the Querying Example. Both the OPTION 1 and the OPTION 2 can be used to describe your phenotype.

In the Step 2 provide the list of the candidate genes that you wish to prioritize. If the Step 2 is ommitted, then the Top 1% of the ranked genome is displayed as a result.

Typical response time is about 3 minutes. If you wish to process your query faster, then check the box for faster processing. The Reset button clears all fields. Click on the Submit button to pass your query for processing.

Example of the query. Score a list of genes for association to cell movement phenotype.

1) **Input the phenotype**: MeSH term for this phenotype is "cell movement". If not sure, then use the link Find Mesh and find most appropriate term for your phenotype. In the MeSH term field the input: cell movement contains **text only** separated by spaces, no special symbols, unless the mesh term has them.

2) **Selecting organism**: Currently we support only mouse, human and fly.

3) **Uploading file**: Upload the gene list, which you would like to score. It should be a simple text file with one column. List the gene identifiers starting at the first row. The gene symbols should be NCBI Gene symbols, not synonyms. We map them into Entrez GeneIDs. The example gene column in the file cell_movement.txt opens in the new window. It contains mouse genes from MGI associated with the cell movement phenotype. You can copy paste this column into a text file on your local computer, and then upload to try this example. **Note**. If gene list is not provided, then the whole genome of the selected organism will be ranked and the top scoring genome genes in the Top 1% of the ranked genme will displayed.

3. Processing

Back to the Table of Contents

The query processing pipeline consists of interrogating our database, sorting the genome of the provided organism based on the information for each data source and creating the report. The completion of every processing step is displayed in the progress page as shown in Figure 2. To open the report with the results click on the provided link **Your results are** here.

Figure 2. Progress Display

Expected processing time of your session is about three minutes

Your query
mesh : Cell_Movement
gofrac : 0.01
homology : 1
dfrac : 0.001
organism : mouse
known genes : Casp1,Cav1,Icam1

Start processing 22:46:43
Literature temporary go_max tables created. 22:47:58
Literature genes and evidence tables created. 22:48:2
Known genes tables created. 22:48:25
DB disconnected. 22:48:26
Finished sorting. 22:48:34
Gene tables created. 22:48:37
Results page created. 22:48:37
Finished . 22:48:37

Your results are here

---

4. List of Prioritized Genes

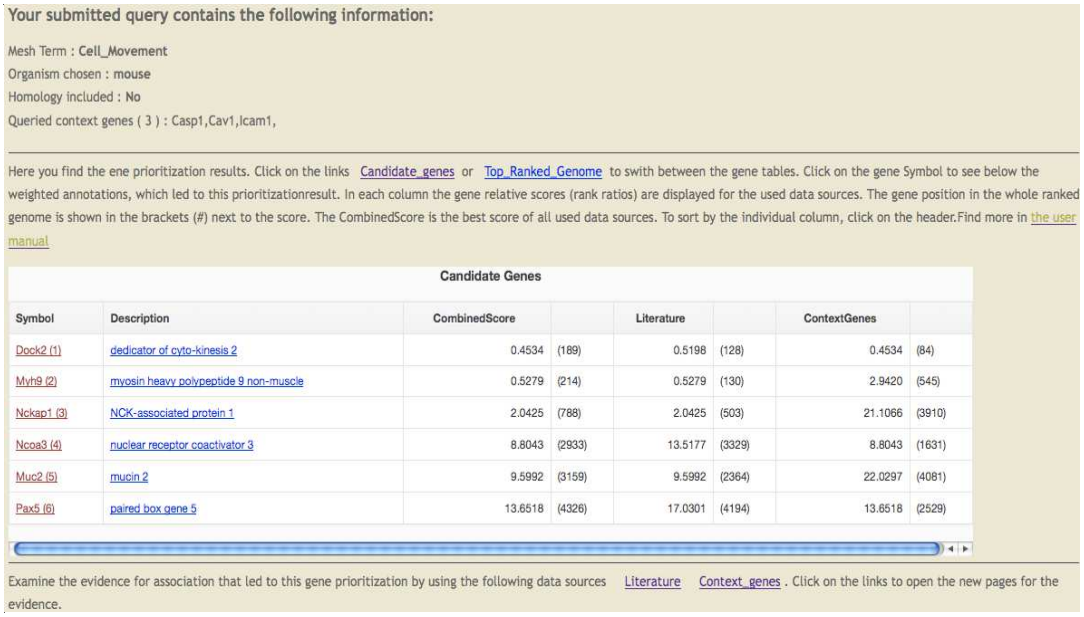
Back to the Table of Contents

The result of the gene prioritization is shown in Figure 3. The list of the candidate genes will be displayed in the prioritization table, if you have submitted the list of the gene candidates. Alternatively the Top 1% of the ranked genome will be shown. The user can switch betwee these views by clicking on the links Candidate Genes or Top Ranked Genome.

The columns in the prioritization table show the ranks of the genes in the whole genome for each selected data source. By clicking on the header of the column, the user can sort the rows by the gene ranks. The gene ranks are assigned by sorting the weighted GO annotations of the gene. The genes having GO annotations with the large weight will be at the top of the ranked list. In the **Literature** column the genes are ranked by the GO weights expressing the strength of the connection between the GO annotation and the MeSH keyword defining the phenotype. In the **Homology+** column the ranking is performed using all weighted GO annotations of the homologous genes. In the **ContextGenes** column the genes are ranked by the GO weights semantically most similar to the GO annotations of the context genes. The details of the weight computation can be found in Technical Report.

The gene ranks are represented by the relative score also termed as a rank ratio. The relative scores (rank ratios) are computed by dividing the gene position in the ranked whole genome by the total number of the ranked genome genes. The relative scores are given a preference to the gene positions in the ranked gene lists. The relative scores allow to compare the gene positions in the ranked lists of the variable size.The smalle relative scores indicate strong relationship between the gene candidate and the phenotype. The gene position in the ranked whole genome is shown in the brackets next to the relative score. The genes in the table are sorted by the **CombinedScore**. The **CombinedScore** is the best (minimum) score out of the all individual scores

Figure 3. Prioritization Results for the Candidate Genes



Your submitted query contains the following information:

Mesh Term : Cell_Movement
Organism chosen : mouse
Homology included : No
Queried context genes (3) : Casp1,Cav1,Icam1,

Here you find the ene prioritization results. Click on the links Candidate_genes or Top_Ranked_Genome to swith between the gene tables. Click on the gene Symbol to see below the weighted annotations, which led to this prioritizationresult. In each column the gene relative scores (rank ratios) are displayed for the used data sources. The gene position in the whole ranked genome is shown in the brackets (#) next to the score. The CombinedScore is the best score of all used data sources. To sort by the individual column, click on the header.Find more in the user manual

Candidate Genes

Symbol	Description	CombinedScore		Literature		ContextGenes	
Dock2 (1)	dedicator of cyto-kinesis 2	0.4534	(189)	0.5198	(128)	0.4534	(84)
Myh9 (2)	myosin heavy polypeptide 9 non-muscle	0.5279	(214)	0.5279	(130)	2.9420	(545)
Nckap1 (3)	NCK-associated protein 1	2.0425	(788)	2.0425	(503)	21.1066	(3910)
Ncoa3 (4)	nuclear receptor coactivator 3	8.8043	(2933)	13.5177	(3329)	8.8043	(1631)
Muc2 (5)	mucin 2	9.5992	(3159)	9.5992	(2364)	22.0297	(4081)
Pax5 (6)	paired box gene 5	13.6518	(4326)	17.0301	(4194)	13.6518	(2529)

Examine the evidence for association that led to this gene prioritization by using the following data sources Literature Context_genes . Click on the links to open the new pages for the evidence.

Click on the gene symbol to see the weighted gene annotations (Figure 4) that led to the current ranking of the genes. The link on the gene description opens the NCBI page of that gene. Examine the reasons for the association between the genes and the phenotypes by clicking on the links Literature and/or the Context Genes below the prioritization table.

By clicking on the gene symbol in the gene prioritization table the user can examine the weighted gene ontology annotations. They are shown in the panel **Weighted GO annotations** in Figure 4. Left table shows the weighted GO annotations that received their weight from the associations of the MeSH Terms in MEDLINE. The right table shows the annotations for which the weights were estimated by the similarity to the functional annotations of the Context Genes. The first row in each table contains the GO annotation with the maximum weight. The maximum weight was the first value, contributing to the obtained rank of the gene.The annotation tables contain detailed information about the sources that contributed to the weights of the annotations.

Contents of the table **Annotations weighted by Literature**:

Column	Description
GO score	the weight of the GO annotation used in the gene ranking
GO term	the annotation description
Organism	the organism that contributed this GO annotation, either the human or mouse or fly
GO-MeSH D Score	the value of the strength of the connection between the GO annotation and the MeSH D term which designates the molecular entity; This value is computed from the frequency of the co-occurrence of the GO annotation and the MeSH D term in the NCBI *gene2go* and *gene2pubmed* databases
MeSH D Term	the name of the MeSH D term with strongest connection to that GO annotation

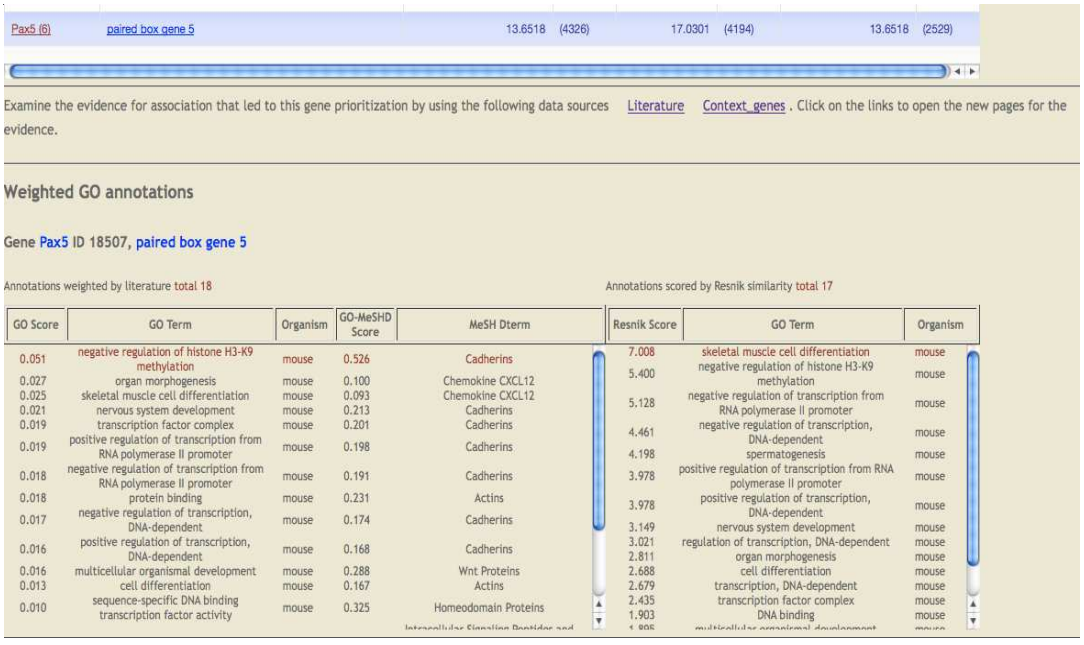
The **GO score** is derived from the **GO-MeSH D Score** and the strength of the connection between the phenotype and that particular **MeSH D Term**. You can explore these relationships between the phenotype and linked **MeSH D Terms** by examining the evidence derived from the literature (link Literature in Figure 3).
The table on the right represents the annotations in which the weights were computed by estimating the semantic similarity between the functional annotations of the **Context Genes** and all other GO annotations in the whole GO database.

Contents of the table **Annotation scored by Resnik similarity**:

Column	Description
Resnik Score	the value of the maximum semantic similarity
GO Term	the name of the GO annotation
Organism	the organism of the gene contributing that annotation

You can explore the GO annotations most contributing to the semantic similarity scores by examining the evidence derived from the Context Genes (link Context Genes in Figure 3).

Figure 4. Display of the weighted gene annotations



5. Evidence derived from Literature

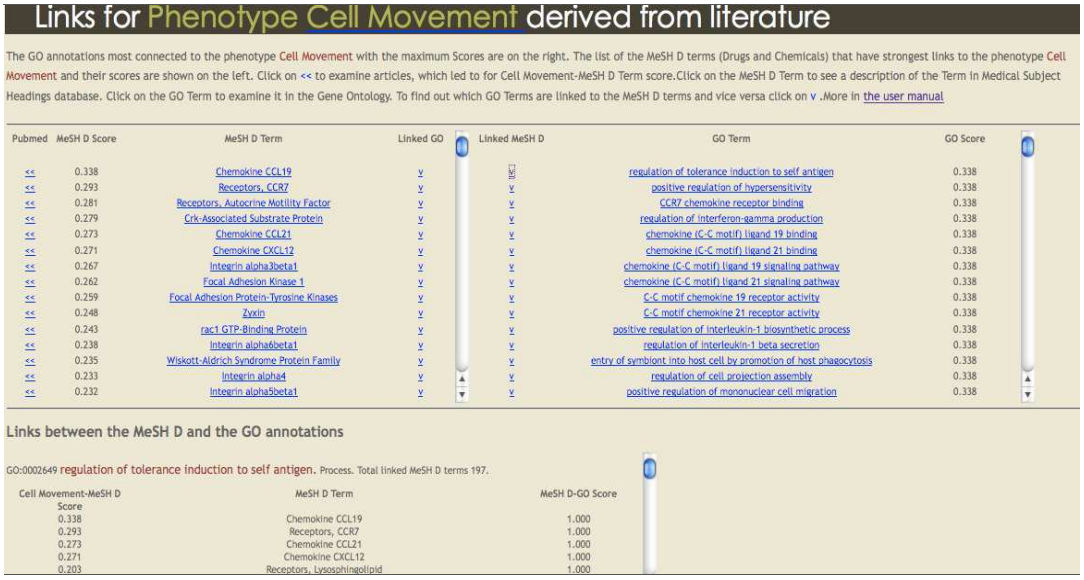
Back to the Table of Contents

The user can examine the relationships between the MeSH terms and GO annotations that led to the current gene prioritization. The evidence page **Links for phenotype derived from literature** is shown in Figure 5. There are two tables and the panel at the bottom. By browsing through the evidence tables the user can see what are the most important links between the concepts, underlying the obtained gene proritization.

The left table shows the MeSH terms of the category D (chemicals and drugs) that have the strongest connections to the phenotype of interest. The strength of these connections is derived from the frequencies of the MeSH annotations occurring together in the MEDLINE articles. The column **MeSH D Score** represents the actual value of the strength of the relationship. The **MeSH D Term** identifies the particular MeSH annotation. By clicking on the MeSH D annotation the user can explore this term in the NCBI MeSH database. The link << opens Pubmed articles, that led to the relationship between the phenotype and that particular MeSH D. The MeSH D terms are linked to the GO annotations. By clicking on v the user will display all GO annotations connected to this particular MeSH D in the bottom panel **Links between the MeSH D and the GO annotation**.

The right table represents the GO annotations most strongly connected to the phenotype through the MeSH D terms. The column **GO Score** shows the actual weight of the GO annotation used in the gene ranking. The GO annotation link opens a new window displaying information in the Gene Ontology database. By clicking on the v in the **Linked MeSH D** column, the user displays all MeSH D terms, associated with that particular GO in the bottom panel **Links between the MeSH D and the GO annotation**. In Figure 5 in the bottom panel the MeSH D terms are listed that were linked to the GO term regulation of tolerance induction to self antigen . The column **Cell Movement-MeSH D** in the bottom table indicates the strength of the connection between the **Cell Movement** phenotype and the MeSH D term. The column **MeSH D-GO Score** shows the strength of the relationship between the regulation of tolerance induction to self antigen and each linked MeSH D term.

Figure 5. Evidence from Literature



Links for Phenotype Cell Movement derived from literature

The GO annotations most connected to the phenotype Cell Movement with the maximum Scores are on the right. The list of the MeSH D terms (Drugs and Chemicals) that have strongest links to the phenotype Cell Movement and their scores are shown on the left. Click on << to examine articles, which led to for Cell Movement-MeSH D Term score.Click on the MeSH D Term to see a description of the Term in Medical Subject Headings database. Click on the GO Term to examine it in the Gene Ontology. To find out which GO Terms are linked to the MeSH D terms and vice versa click on v .More in the user manual

Pubmed	MeSH D Score	MeSH D Term	Linked GO	Linked MeSH D	GO Term	GO Score
<<	0.338	Chemokine CCL19	v	v	regulation of tolerance induction to self antigen	0.338
<<	0.293	Receptors, CCR7	v	v	positive regulation of hypersensitivity	0.338
<<	0.281	Receptors, Autocrine Motility Factor	v	v	CCR7 chemokine receptor binding	0.338
<<	0.279	Crk-Associated Substrate Protein	v	v	regulation of interferon-gamma production	0.338
<<	0.273	Chemokine CCL21	v	v	chemokine (C-C motif) ligand 19 binding	0.338
<<	0.271	Chemokine CXCL12	v	v	chemokine (C-C motif) ligand 21 binding	0.338
<<	0.267	Integrin alpha3beta1	v	v	chemokine (C-C motif) ligand 19 signaling pathway	0.338
<<	0.262	Focal Adhesion Kinase 1	v	v	chemokine (C-C motif) ligand 21 signaling pathway	0.338
<<	0.259	Focal Adhesion Protein-Tyrosine Kinases	v	v	C-C motif chemokine 19 receptor activity	0.338
<<	0.248	Zyxin	v	v	C-C motif chemokine 21 receptor activity	0.338
<<	0.243	rac1 GTP-Binding Protein	v	v	positive regulation of interleukin-1 biosynthetic process	0.338
<<	0.238	Integrin alpha6beta1	v	v	regulation of interleukin-1 beta secretion	0.338
<<	0.235	Wiskott-Aldrich Syndrome Protein Family	v	v	entry of symbiont into host cell by promotion of host phagocytosis	0.338
<<	0.233	Integrin alpha4	v	v	regulation of cell projection assembly	0.338
<<	0.232	Integrin alpha5beta1	v	v	positive regulation of mononuclear cell migration	0.338

Links between the MeSH D and the GO annotations

GO:0002649 regulation of tolerance induction to self antigen. Process. Total linked MeSH D terms 197.

Cell Movement-MeSH D Score	MeSH D Term	MeSH D-GO Score
0.338	Chemokine CCL19	1.000
0.293	Receptors, CCR7	1.000
0.273	Chemokine CCL21	1.000
0.271	Chemokine CXCL12	1.000
0.203	Receptors, Lysosphingolipid	1.000

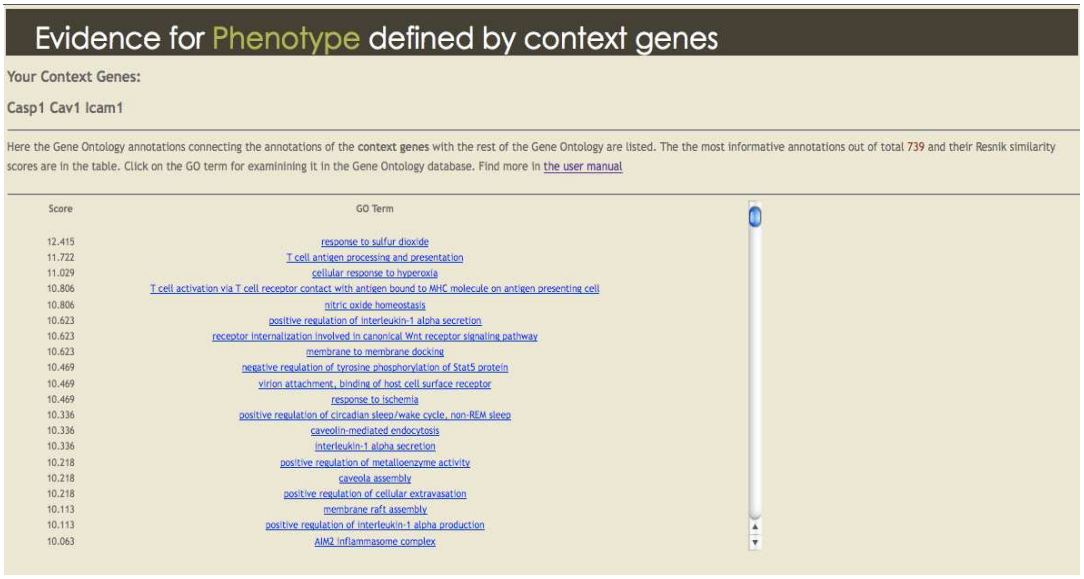
6. Evidence derived from Context Genes

Back to the Table of Contents

The page **Evidence for Phenotype defined by context genes** in Figure 6 shows the Gene Ontology annotations that provided semantic similarity scores for other GO terms. The listed GO annotations represent the most informative common ancestors in the taxonmy of the Gene Ontlogy. The most informative common ancestors relate the annotations of the **Context Genes** and all ather GO annotations in the whole database. The context genes defining the phenotype are shown in the top panel.

The **Score** column indicates the information content (IC) assigned to the GO annotation. The IC is computed as a negative logarithm of the ratio of the number of genes annotated by that GO term and all it's descendants to the number of all annotated genes. If the GO term is very specifc, then it annotates just few genes and has very high information content. High information content of the annotation indicates that alldescendants of this annotation in the taxonomy will be very informative as well. They all will have high GO **Scores**, which are used to rank the genome genes. The genes very similar functionally to the context genes will be ranked very high.

Figure 6. Evidence from Context Genes



Evidence for Phenotype defined by context genes

Your Context Genes:

Casp1 Cav1 Icam1

Here the Gene Ontology annotations connecting the annotations of the **context genes** with the rest of the Gene Ontology are listed. The the most informative annotations out of total 739 and their Resnik similarity scores are in the table. Click on the GO term for examininng it in the Gene Ontology database. Find more in the user manual

Score	GO Term
12.415	response to sulfur dioxide
11.722	T cell antigen processing and presentation
11.029	cellular response to hyperoxia
10.806	T cell activation via T cell receptor contact with antigen bound to MHC molecule on antigen presenting cell
10.806	nitric oxide homeostasis
10.623	positive regulation of interleukin-1 alpha secretion
10.623	receptor internalization involved in canonical Wnt receptor signaling pathway
10.623	membrane to membrane docking
10.469	negative regulation of tyrosine phosphorylation of Stat5 protein
10.469	virion attachment, binding of host cell surface receptor
10.469	response to ischemia
10.336	positive regulation of circadian sleep/wake cycle, non-REM sleep
10.336	caveolin-mediated endocytosis
10.336	interleukin-1 alpha secretion
10.218	positive regulation of metalloenzyme activity
10.218	caveola assembly
10.218	positive regulation of cellular extravasation
10.113	membrane raft assembly
10.113	positive regulation of interleukin-1 alpha production
10.063	AIM2 inflammasome complex

Back to the Table of Contents

Appendix B. Supplementary Material for Chapter 2

In this Appendix the details about the created pipeline tools are provided. The tools as a tarball have been made available at phenotypelinks.ogic.ca/chip-seq-pipeline.html. The interfaces of the tools to the Galaxy workflow management system will be available from figshare.com/PhenotypeLinks.

B.10 Dependencies of the ChIP-Seq Data Analysis Tools

The current dependencies of the tools are summarized in Table B.10.

TABLE B.10: Programs and Data Required for the Implemented ChIP-Seq Data Analysis Pipeline Tools on Unix-like System

Description	**Link to the Source**
Human genome hg19 GRCh37 build	The human genome build released in February 2009 by Genome Reference Consortium GRCh37 (`genome.ucsc.edu`)
RefSeq Gene table	downloaded as of June 2013 from the UCSC Table browser (`genome.ucsc.edu`)
Python 2.6 version or higher	`www.python.org/download/releases/`
Perl 5.0 version or higher	`www.perl.org/get.html`
BedTools 17.0 version	`code.google.com/p/bedtools/downloads/list`
Sequence Alignment Map tools SAMtools	`samtools.sourceforge.net/`
Mapping and assembly with scores program MAQ	`maq.sourceforge.net/`
Burrows Wheeler Aligner program BWA	`bio-bwa.sourceforge.net/`
Model based analysis for ChIP-Seq MACS	`liulab.dfci.harvard.edu/MACS/`
R 2.15 or higher	`www.r-project.org/`
UCSC utilities to work with *bed* and *wig* file formats	`http://hgdownload.cse.ucsc.edu/admin/exe/`
R stringr package	available for installation through R function install.packages()
Bioconductor	`www.bioconductor.org/install/`
Bioconductor packages Biostrings, motifRG, SeqLogo	available for installation through Bioconductor function biocLite()
Galaxy workflow management system	`usegalaxy.org`

B.11 Characterization Genome Wide Binding of the Transcription Factor

To assign the genomic features to the peaks and compute the summary three programs were created and integrated into Galaxy: `Closest Bed`, `Peak Genomic Features` and `Peak Counts over Genomic Features`. The assignment of the genomic regions to the peaks is achieved by several computational steps. The workflow of peak association with the genomic features is presented in Figure B.11. There the arrows show the order of operations. The genomic coordinates of all documented gene accessions are extracted from the RefSeq UCSC gene table. In Step 1 for each peak the closest or intersecting gene regions are identified by the `Closest Bed` tool. The output of the Step 1 is described in Panel Step1 of Table B.11. In Step 2 each genomic feature intersecting the peak is identified and recorded by using the `Peak Genomic Features` tool. The peak region can intersect multiple genomic features: the promoter, the 5' and 3' untranslated regions, the exons, the introns and the intergenic area. The output of the Step 2 is described in Panel Step2 in Table B.11.

In Step 3 each peak is associated only with one genomic feature by using the `Peak counts over Genomic Features` tool. The result of the Step 3 is the summary table of the overall association of genomic features with peaks. The genomic features are associated with a peak based on the precedence: promoter, 3' UTR, 5' UTR, exon, intron and inter-genic area. For example, if a peak intersects an inter-genic region, and a promoter, and 5' UTR then it will be associated with the promoter. If a peak intersects an exon and intron then it will be associated with only the exon.

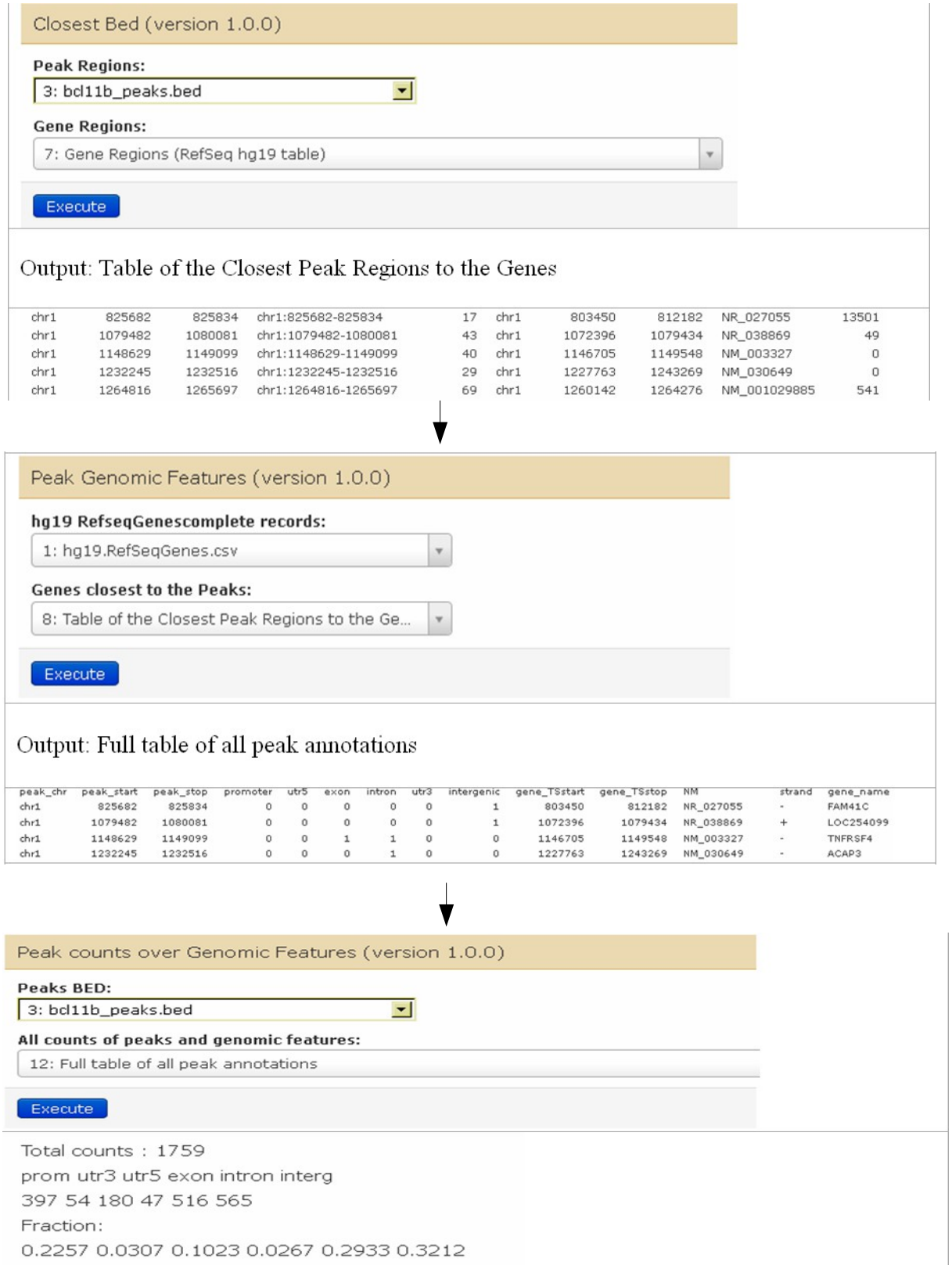


FIGURE B.11: Computation of Genomic Features Occupancy by Peaks. User Interface to the Tools

Table B.11: Outputs of Step1 and Step2

Table B.11: Step1. Output of the `Closest Bed` Tool

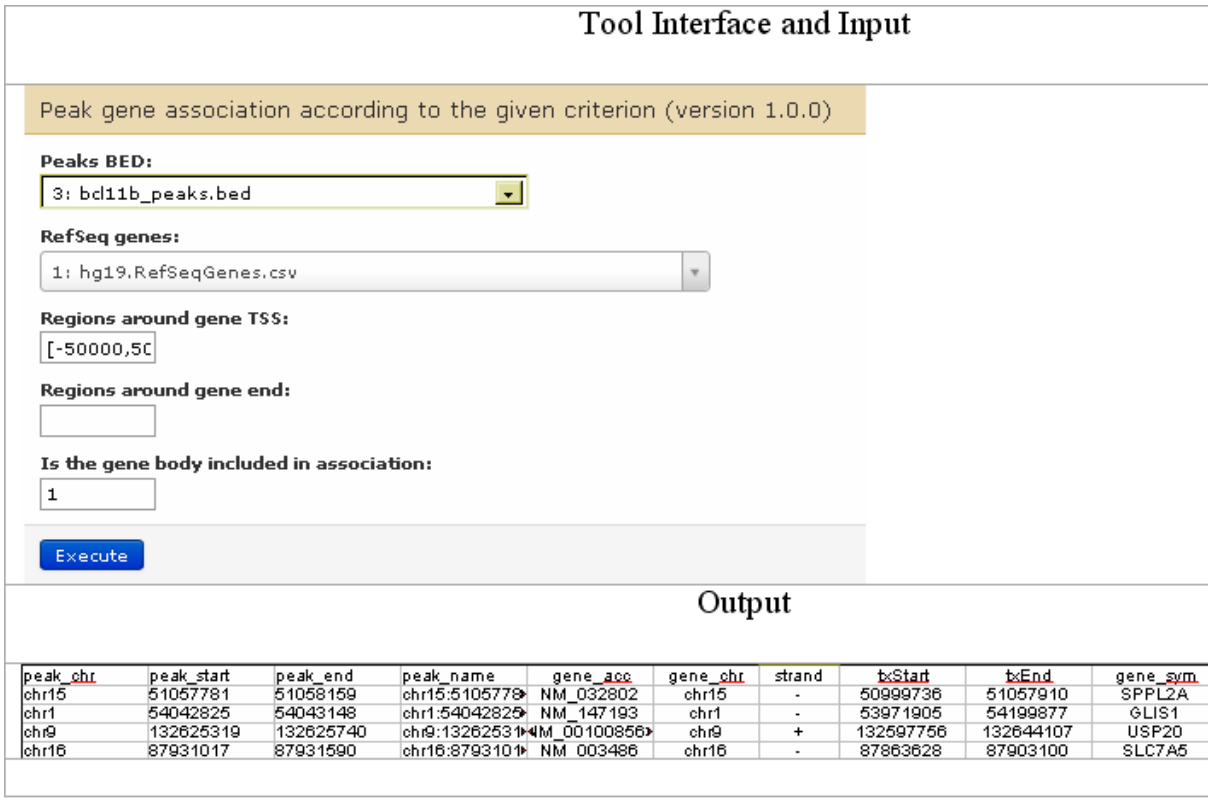
Fields	**Values**
Peak Chromosome	chr1
Peak Chromosome start	1079482
Peak Chromosome end	1080081
Peak name	chr1:1079482-1080081
Number of tags in peak	43
Gene Chromosome	chr1
Gene Chromosome start	1072369
Gene Chromosome end	1079434
Gene accession number	NR_038869
Absolute distance+1 in bp between the closest coordinates of the peak and gene	49

Table B.11: Step2. Output of the `Peak Genomic Features` Tool

Fields	**Values**
Peak Chromosome	chr1
Peak Chromosome start coordinate	1148629
Peak Chromosome end coordinate	1149099
Promoter flag	0
Utr 5' flag	0
Exon flag	1
Intron flag	1
Utr 3' flag	0
Intergenic flag	0
Gene TSS start coordinate	1146705
Gene end coordinate	1149548
Gene accession number	NM_003327
Gene strand	−
Gene symbol	TNFRSF4

B.12 Association of Peaks to Genes

The interface and output of the tool to associate the genes with the peaks are shown in Figure B.12. The required inputs consist of a RefSeq gene table from the UCSC Table Browser, the list of the genomic regions of peaks (in BED format) and the desired distance criteria. The tool computes the regions around each gene called query regions. The genes for which the query regions intersect the peaks are associated with those peaks. The distance can de defined related



peak_chr	peak_start	peak_end	peak_name	gene_acc	gene_chr	strand	txStart	txEnd	gene_sym
chr15	51057781	51058159	chr15:5105778	NM_032802	chr15	-	50999736	51057910	SPPL2A
chr1	54042825	54043148	chr1:54042825	NM_147193	chr1	-	53971905	54199877	GLIS1
chr9	132625319	132625740	chr9:13262531	NM_001008566	chr9	+	132597756	132644107	USP20
chr16	87931017	87931590	chr16:8793101	NM_003486	chr16	-	87863628	87903100	SLC7A5

FIGURE B.12: Interface to the Tool `Gene to Peak Association According to Various Distance Criteria`

to the TSS or the full length of the gene. The specific query regions of various size and separated from the gene TSS by various distances can be used to associate the genes with the peaks. In the intermediate processing step the tool uses the `intersectBed` command of the BedTools [Quinlan and Hall, 2010] producing list of genes associated with the given peaks. The output of the tool

is explained in Table B.12. The output of this tool can be used for other related analysis for example analyzing the intersections of several gene lists.

TABLE B.12: Output of the Tool `Peak gene association according to the given criterion`

Fields	**Values**
Peak Chromosome	chr16
Peak Chromosome start coordinate	87931017
Peak Chromosome end coordinate	87931590
Peak name	chr16:87931017-87931590
Accession of the closest associated gene	NM_004386
Gene Chromosome	chr16
Gene strand	–
Gene TSS start coordinate	87863628
Gene end coordinate	87903100
Gene symbol	SLC7A5
Query start coordinate	87853100
Query end coordinate	87953100

B.13 Motif Analysis in Peaks

The motif tools in the current pipeline are used to perform four tasks: extraction of sequences in FASTA format for the given genomic regions (the `Fasta from Bed` tool), randomization of DNA sequences (the `ishuffle` tool), finding all occurrences of a particular pattern in the peak sequences (the `seq_in_peaks` tool) and *de novo* pattern search in peaks by motifRG (the `motifs_denovo` tool). The tools use motifs specified by consensus sequence using IUPAC code [Morgan et al., 2009].

B.14 The `Fasta from Bed` Tool

The tool `Fasta from Bed` extracts the DNA sequence of the peaks by executing the `FastaFromBed` command of the BedTools. The tool requires only one input the list of the genomic regions and its interface is shown in Figure B.14.

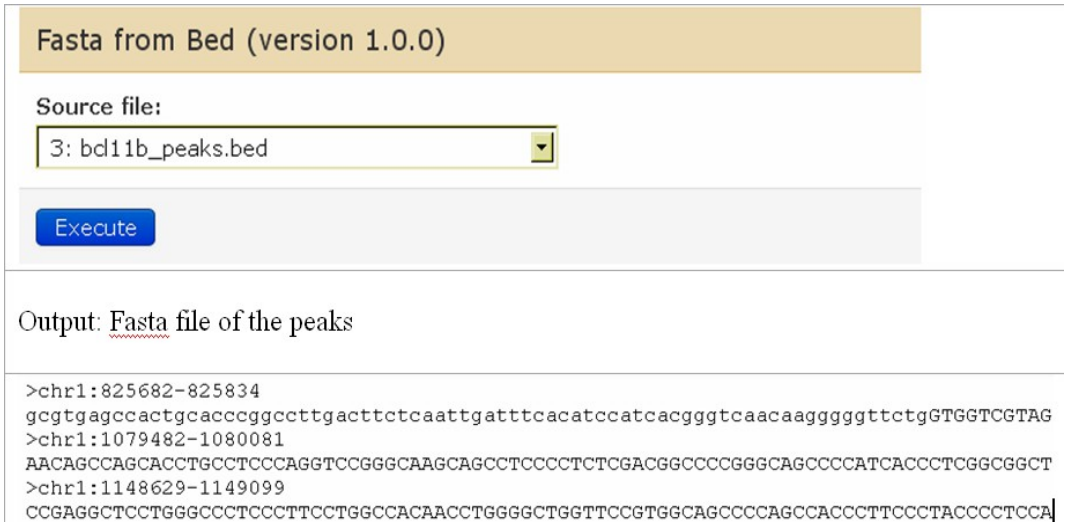


FIGURE B.14: Interface to the Tool `Fasta from Bed`

B.15 The `ishuffle` Tool

The `ishuffle` tool randomizes the provided genomic sequences to create the background sets for the motif discovery. This tool preserves the original frequencies of dinucleotides. The `ishuffle` tool is a wrapper of the open source program ushuffle [Jiang et al., 2008] which is integrated into the pipeline tool set. The interface to this tool is shown in Figure B.15. At input the tool requires the FASTA file, the number of randomizations, the size of dinucleotide (K-mer) to preserve the frequency of occurrences in the randomized sequence and a random seed.

B.16 The `seq_in_peaks` Tool

The tool `seq_in_peaks` finds all occurrences of a motif pattern given as a consensus sequence in the DNA sequences of the genomic regions. The search is performed using the forward and reverse version of the pattern. The interface to this tool is shown in Figure B.16. At input the tool requires the motif given by consensus sequence and the

FIGURE B.15: Interface to the Tool `ishuffle` Randomizing the DNA Sequences

le of the DNA sequences in which to perform the search. The output of the tool consists of the table listing all matches to the given pattern. Each row of the output table contains the peak coordinates, list of the exact pattern matches, the coordinates of the matches in the peak and the types of the matches shown in Table B.16.

B.17 The *motifs_denovo* Tool

The *de novo* motif identification tool is a wrapper of motifRg [Fong et al., 2012] Bioconductor package. It requires background sequences in order to find motifs enriched in the peak sequences. The background can consist of the sequences flanking the peak regions, randomly shuffled peak sequences, or an entire Chromosome of the genome. The motifRG package uses the expectation maximization algorithm to identify the enriched motif sequences in the subset of the DNA FASTA sequences of the peaks. The motif sequences enriched in the foreground (peaks) are identified by the value TRUE in the third column of the output. The identified motifs might vary in different runs using different background sequences. However, this limitation is present in all tools that use background DNA sequences to find the over-represented motifs. The interface to this tool

TABLE B.16: Output of the `seq_in_peaks` Tool

Fields	**Values**
Peak coordinates	chr1:2470002-2470605
Matching Patterns	CCCCTG GTCCCC
Coordinates of the matching patterns in peaks	156 161, 166 171
Type of the matching	ff cf

is shown in Figure B.17. The tool requires to provide the FASTA files of the genomic regions in which the *de novo* search is performed, the background FASTA sequences and the desired number of motifs to be identified. The tool has a lengthy execution time for large number of the motifs (10 and above). The output of the *motifs_denovo* tool is a table of motifs, their scores and number of occurrences. The output fields are explained in Table B.17.

B.18 Analysis of Transcription Factor Composite Sequences

The `pref_dist_in_peaks` tool was implemented in R. The interface to this tool is shown in Figure B.18. At input it requires two motif sequences. One serves as an anchor motif. The second motif will have varying locations with respect to the anchor motif within the range of the predefined distances upstream or downstream of the anchor motif. A list of the composite motifs in which the distances between the anchor motif and the second motif vary within the range of 20 base pairs is created. An exact pattern matching of the created composite motifs is performed in the DNA FASTA sequences of the peaks and in Chromosome 1. The Chromosome masked for repeats was chosen as background to use with the `pref_dist_in_peaks` tool.

There are four types of the identified matches. The peak sequences are tested in the original and complementary form. The composite motifs are matched to those sequences in both the forward and the reverse orientation. The codes of the orientation are as follows: o_lr original peak sequence , the composite is forward (left to right); o_rl original peak sequence , the composite is reversed (right to left); c_lr complement peak sequence , the composite is forward (left to

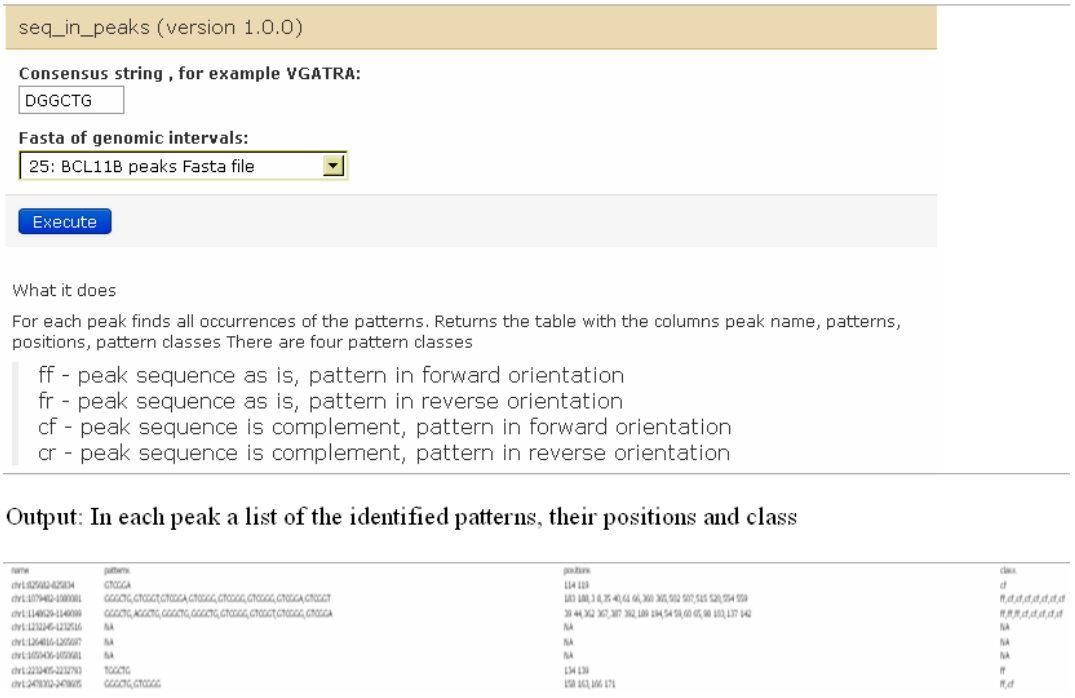


FIGURE B.16: Interface to the Tool `seq_in_peaks`

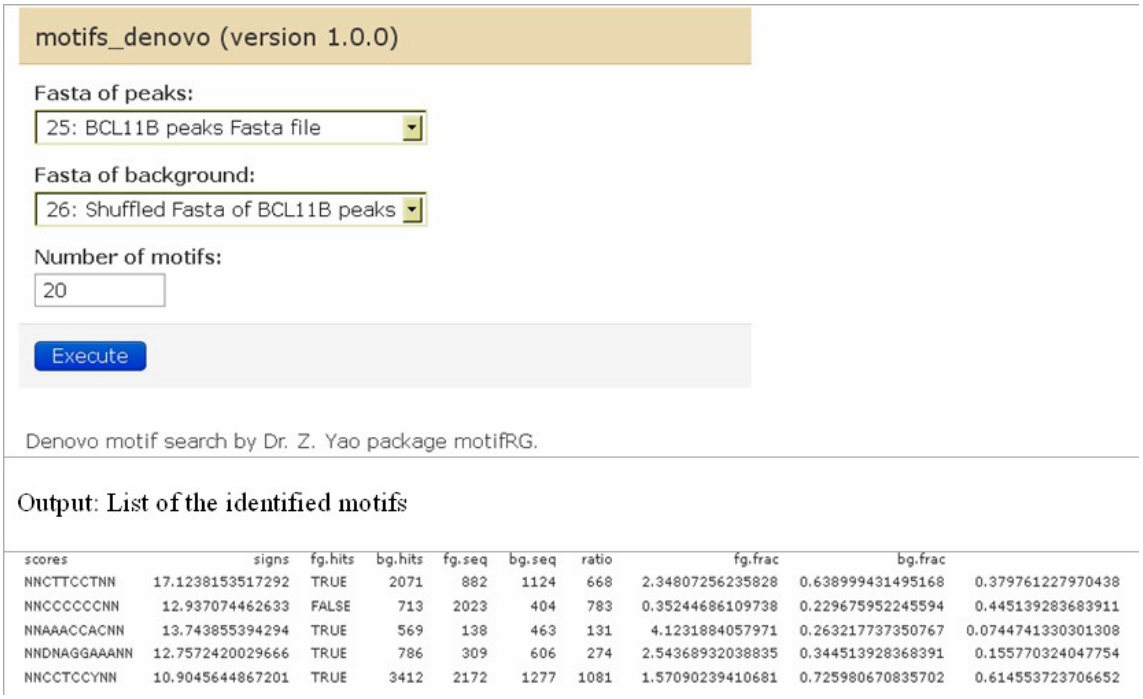
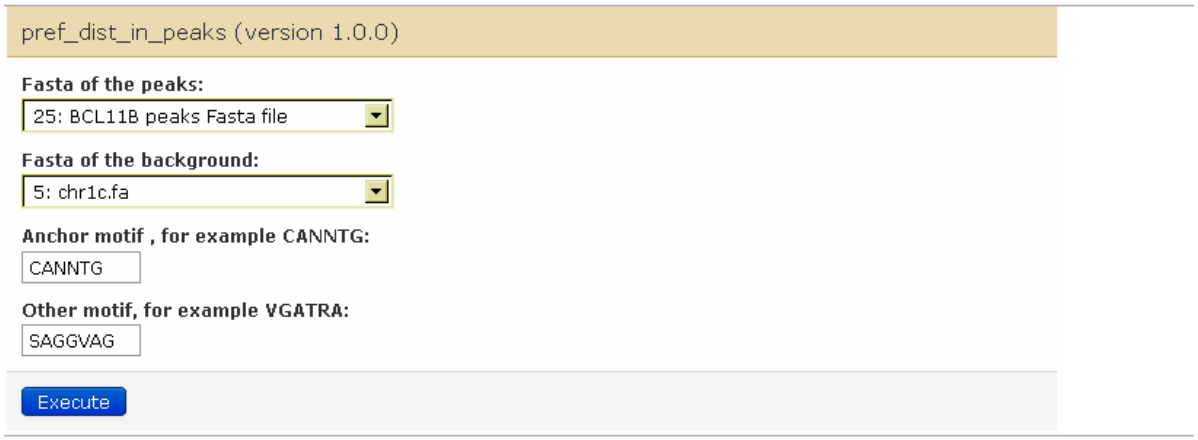


FIGURE B.17: Interface to the `motifs_denovo` Tool to Perform the *de novo* Motif Search

TABLE B.17: Output of the `motifs_denovo` Tool

Fields	**Values**
Consensus sequence of the motif	NNCTTCCTNN
Score for the identified motif	17.1238
Motif in foreground	TRUE
Number of hits in foreground	2071
Number of hits in background	882
Number of foreground sequences tested	1124
Number of background sequences tested	668
Ratio of foreground hits to background hits	2.348
Fraction of hits in foreground	0.6389
Fraction of hits in background	0.3797

right); c_rl complement peak sequence , the composite is reversed (right to left). The output table contains all matches of the composite motifs found in the peak DNA sequences. The same peak can have several matching composites. All matches are used to compute the frequencies of the observed distances between the anchor and the second motif. The relative motif orientation is accounted for by the user providing the required sequence orientation in specification of the anchor and other motif. Table B.18 explains the output. The frequencies at which the composite with the particular distance between the motif occur are normalized by the total number of matches. A plot which displays the frequencies of the occurrences of the composite motifs at the distances upstream [-20..0]bp and downstream [0..20]bp of the anchor motif in the peaks and the Chromosome1 is created. An example of the plot of the preferred distances is in Figure 4.4 in Section 4.1.



Output:

peak_no	peak_name	orient	mstart	mstop	peak_len	dist	pattern
2	chr1:1079482-1080081	o_lr	120	137	599	5	CA[A,C,G,T][A,C,G,T]TG.....[C,G]AGG[A,C,G]AG
5	chr1:1264816-1265697	c_rl	13	43	881	18	CA[A,C,G,T][A,C,G,T]TG.....[C,G]AGG[A,C,G]AG
8	chr1:2478302-2478605	o_rl	179	197	303	6	CA[A,C,G,T][A,C,G,T]TG.....[C,G]AGG[A,C,G]AG
11	chr1:6035583-6035909	o_lr	196	217	326	-9	[C,G]AGG[A,C,G]AG.....CA[A,C,G,T][A,C,G,T]TG
11	chr1:6035583-6035909	o_lr	176	202	326	14	CA[A,C,G,T][A,C,G,T]TG.....[C,G]AGG[A,C,G]AG
12	chr1:6086103-6086503	c_rl	79	102	400	-11	[C,G]AGG[A,C,G]AG.....CA[A,C,G,T][A,C,G,T]TG
12	chr1:6086103-6086503	c_rl	79	91	400	0	[C,G]AGG[A,C,G]AGCA[A,C,G,T][A,C,G,T]TG
12	chr1:6086103-6086503	c_rl	96	118	400	10	CA[A,C,G,T][A,C,G,T]TG.....[C,G]AGG[A,C,G]AG

FIGURE B.18: Interface to the `pref_dist_in_peaks` Tool to Identify Preferred Distances in the Composite Motifs

TABLE B.18: Output of the pref_dist_in_peaks Tool

Fields	**Values**
Peak number	2
Peak name	chr1:1079482-1080081
Orientation of the match of the composite sequence	o_lr
Start coordinate of the match in peak	120
End coordinate of the match in peak	137
Length of the peak	599
Distance between the motifs in the composite sequence	5
Pattern of the composite motif	CA[A,C,G,T][A,C,G,T]TG .....[C,G]AGG[A,C,G]AG

B.19 Program to Remove the Bad Quality Reads

```
#####################################################
#  removeB.pl
#
#  Input    FASTQ file at the standard input
#  Output   FASTQ at the standard output
#           sequences with BBB stretches removed
#
#  Call     $cat sequences.FASTQ | perl removeB.pl > sequences_filtered.FASTQ
#####################################################
#!usr/bin/perl
use strict;
my $runcount=0;
my $total=0;
my @fourlines;
my $filtered=0;
while (<> )
{$runcount=$runcount+1;
  push @fourlines,$_;
  if ( $runcount==4)
  {
   my $line4 = pop @fourlines;
   my $line3 = pop @fourlines;
   my $line2 = pop @fourlines;
   my $line1 = pop @fourlines;
   $runcount=0;
   $total=$total+1;
   if ( index($line4 ,"BBBBBBBBBB") == -1)
   {
     print join("",$line1, $line2,$line3,$line4);
     $filtered=$filtered+1;
   }
  }
}
system("echo Total $total, Filtered $filtered");
exit;
```

Listing 1: Removal of Bad Quality Reads from FASTQ file

B.20 Commands to Prepare to Run in Parallel the Alignment by MAQ

```
#transform the clean short reads FASTQ file into sanger FASTQ format

>maq sol2sanger sequences.FASTQ sequences_sanger.FASTQ

#make the reference genome ready for use with MAQ.
#The reference hg19 genome in 2bit format downloaded from UCSC
#is transformed into the fasta format. The chromosomes in chrList
#1-22 and X and M and Y are included

>twoBitToFa -seqList=chrList hg19.2bit hg19.fa

#the files have to be transformed into binary MAQ format.
#The reference genome is transformed

>maq fasta2bfa hg19.fa hg19.bfa

# Short read file is split into smaller files of 4 million reads.
# The split files should be converted into binary bfq format.
# The split files are in the directory BFQ.
# The prefix for split files is bcl_.

>split --lines=4000000 sequences_sanger.FASTQ BFQ/bcl_

# List the full paths
>ls -f `pwd`/BFQ/bcl_* > file_list

#Variable contains how many files were created
>NFILES=`wc --lines < file_list`

# Submitting the alignment tasks to Sun Grid Engine
# make separate directory for the results
>mkdir RESULTS

#The qsub program submits the tasks to the queue for the execution.
#The required parameters are the error and output directories,
#number of tasks to run and the name of the shell script
#to be executed on each separate computing cluster node.

>qsub -e `pwd`/RESULTS -o `pwd`/RESULTS -t 1-$(( $NFILES)) MAQaln.sh

# After the alignment
# the files containing aligned reads should be merged

>maq mapmerge BCL11B.map BFQ/*.bfq.map

#unmapped reads are in text files *.unmap
>cat BFQ/*.bfq.unmap >BCL11B.unmapped.txt

# The reads that map to several locations are nonunique
# The nonunique reads are filtered by the filtermap.pl script
# This script only checks the mapping quality
# If mapping quality is zero the read is directed to the
# file of non unique reads otherwise into the unique reads file

>maq mapview BCL11B.map | perl ../TOOLS/filtermap.pl

# The mapping resulted in two files
# nonunique.aln.txt and unique.aln.txt
```

Listing 2: Preparatory Steps for the Sequence Alignment with MAQ

B.21 Code to Submit the Alignment Tasks to Run on the Computer Cluster by Using Sun Grid Engine

```
#####################################
# MAQaln.sh
#####################################
#!/bin/ssh
# request Bourne shell as shell for job
#$ -S /bin/sh
#$ -cwd
#$ -q all.q
#

cd RESULTS

# date and time
date

# retrieve file name corresponding to the ID
# of the task running on the current core
# from the list of the files that have to be aligned
#
#the alignment files are identified through the environment
#variable SGE_TASK_ID which tells the core the identifier
#of the task the core is running

INFILE=`awk "NR== $SGE_TASK_ID" ../file_list`

# transform file into binary MAQ format

maq FASTQ2bfq $INFILE $INFILE.$SGE_TASK_ID

# mapping program produces map and unmap file for every input file
# each task is executed independently of the others
# the original sequence data are in file $INFILE.$SGE_TASK_ID

maq match -u $INFILE.$SGE_TASK_ID.unmap $INFILE.$SGE_TASK_ID.map hg19.bfa
    $INFILE.$SGE_TASK_ID

# print date and time again

date
```

Listing 3: Submission of Tasks to the Computing Cluster Using SGE

B.22 Commands to Prepare the Alignment File to Visualize in the UCSC Genome Browser

```
#The awk program is used for manipulation of the columns of aligned files
# and perform the extension of reads in sequencing direction using the leftmost
# $5$' end of the read as anchor.
#
#The following line of code shows commands to create the BED file
#from the unique.aln.txt file in which the reads are 50bp long,
#to remove the reads aligned to mitochondrial Chromosome chrM,
#to change the space separator into TAB separator and sorts the BED file.
#The resulting file is aln.bed
#
>grep ^chrM -v BCL11B.aln.bed >t; tr " " "\t" < t > tt; sortBed -i tt >aln.bed
#
#The following the aln.ext.bed in which the reads are already extended to their
#original length
#calls the BedTools command genomeCoverageBed to create the GenomeGraph file
#
>genomeCoverageBed -bga -i aln.ext.bed -g hg19.genome > aln.bedGraph
#
#The GenomeGraph file may be very large.
#It is transformed into into the BigWig required by UCSC genome browser
#by the BedTools command bedGraphToBigWig
#
>bedGraphToBigWig aln.bedGraph hg19.genome aln.bw
#
#The resulting aln.bw file is uploaded to the UCSC browser for visualization.
```

Listing 4: Prepare the Alignment to Visualize in the UCSC Genome Browser

B.23 Commands to Call MACS

```
# The columns in the BED files used by MACS should be separated by spaces
# Change the tabs into the spaces

>tr "\t" " " < control.aln.bed> t; mv t control.aln.bed
>tr "\t" " " < signal.aln.bed> t; mv t signal.aln.bed

# Call MACS with parameters specified in the BCL11B analysis
# Provide the aligned BED files for the signal and the control
>macs -t signal.aln.bed -c control.aln.bed --name=BCL11B --tsize=50 --mfold=10
    --bw=200 --pvalue=1e-7

# The coverage analysis is performed by
# fisher_yates_shuffle_withMock.pl script
# using MACS with the set of the final parameters

>perl fisher_yates_shuffle_withMock.pl signal.aln.bed control.aln.bed
```

Listing 5: Commands to Call the MACS Program

B.24 Perl Script to Compute the Coverage of the Peaks by the Short Reads

```
##############################################################################
## Read coverage with Mock - using the modified fisher_yates_shuffle.pl
## fisher_yates_shuffle_withMock.pl
## Shuffling is performed to both signal and mock file
##############################################
# compute read coverage
#
# fisher_yates_shuffle_withMock.pl
# Input:   The alignment bed file
# Output:  The statistics table at STDOUT
#          The same table in file stats
#
# Call:    perl fisher_yates_shuffle_withMock.pl signal.aln.bed control.aln.bed
##############################################
# compute read coverage
# fisher_yates_shuffle( \@array ) : generate a random permutation # of @array in place
##############################################

use strict;

my $file_sig = shift;
my $file_mock = shift;

my @array_sig;
my @reads_sig;
my %readID_sig;

my @array_mock;
my @reads_mock;
my %readID_mock;

my $i;
my $status=0;

#system("/data/binaries/BEDTools-Version2.8/bin/intersectBed -wb -a BCL11B_peaks.bed
    -b BCL11B.aln.bed | awk \'{print $9}\' ");

# processing signal
open( FILE, "<$file_sig");while (<FILE>){ push(@reads_sig,$_);

my ($chr,$start,$end,$id,$q,$strand)=split("\t",$_);
$readID_sig{$id}=$status; }
close(FILE);print "Finished reading signal\n";

#processing mock
open( FILE, "<$file_mock");while (<FILE>){ push(@reads_mock,$_);

my ($chr,$start,$end,$id,$q,$strand)=split("\t",$_);
$readID_mock{$id}=$status; }
close(FILE);print "Finished reading mock\n";

my $n_sig=scalar @reads_sig;
for ( $i=0; $i<$n_sig;$i+=1){ push (@array_sig,$i);}
print "Finished creating signal n=",$n_sig,"\n";

my $n_mock=scalar @reads_mock;
for ( $i=0; $i<$n_mock;$i+=1){ push (@array_mock,$i);}
```

```
print "Finished creating mock n=",$n_mock,"\n";

fisher_yates_shuffle( \@array_sig ); # permutes @array in place
print "Finished Shuffling of signal\n";
fisher_yates_shuffle( \@array_mock ); # permutes @array in place
print "Finished Shuffling of mock\n";

my @ranges_sig;
my $step_sig=int( $n_sig/20);

# create from_to indexes
for($i=$step_sig; $i<$n_sig; $i+=$step_sig ){ push (@ranges_sig,$i);}
push(@ranges_sig, $n_sig);
print "Signal ranges \n";
print join(" ",@ranges_sig);

my @ranges_mock;
my $step_mock=int( $n_mock/20);
# create from_to indexes
for($i=$step_mock; $i<$n_mock; $i+=$step_mock ){ push (@ranges_mock,$i);}
push(@ranges_mock, $n_mock);
print "Mock ranges \n";
print join(" ",@ranges_mock);

#########################################################################
#  adding more reads in pieces
#  print the shuffled reads
#  sort them
#  call MACS peaks
#  intersect peaks and reads with intersectBed
#  input the reads that intersect with peaks
#  count the reads that changed status from background to foreground
#########################################################################

my @statistics;
my $i;

foreach my $limit_signal(@ranges_sig)
{
        my $limit_mock=shift @ranges_mock;
        open(OUTFILE_SIG,">out_signal");
        open(OUTFILE_MOCK,">out_mock");

        for ( $i=0; $i<=$limit_signal; $i+=1){
        # take index from shuffled array
        my $si=$array_sig[$i];
        print OUTFILE_SIG join(\t,split( ,$reads_sig[$si])),\n;}
        print Created  out_signal $limit_signal \n;
        for ( $i=0; $i<=$limit_mock; $i+=1){
        # take index from shuffled array
        my $si=$array_mock[$i];
        print OUTFILE_MOCK join(\t,split( ,$reads_mock[$si])),\n;}
  print Createdoutmock$limit_mock \n;
        close(OUTFILE_SIG);
        close(OUTFILE_MOCK);

#### we created the files for peak calling
#### lets sort them  The files should be tab delimited
        system("/data/binaries/BEDTools-Version2.8/bin/sortBed -i out_signal >
    out_signal.bed");
        print "Sorted shuffled out_signal.bed \n";

        system("/data/binaries/BEDTools-Version2.8/bin/sortBed -i out_mock >
    out_mock.bed");
```

```
        print "Sorted shuffled out_mock.bed \n";

#### lets call peaks
####  parameters - lambda estimated from background, fold change =4, bandwidth=180/2
###   parameters - adjusted for peak calling process

system("/data/binaries/MACS-1.3.7.1/bin/macs -t out_signal.bed -c out_mock.bed
    --name=out --tsize=50 --mfold=10 --bw=200");
print "Peaks called\n";

#### lets intersect the peak file with reads file , only for tab delimited files : tr
    " " \t<BED/BCL11B.aln.bed > t_sig
        print "/data/binaries/BEDTools-Version2.8/bin/intersectBed -wb -a
    out_peaks.bed -b t_sig  > out.reads \n";
        system("/data/binaries/BEDTools-Version2.8/bin/intersectBed -wb -a
    out_peaks.bed -b t_sig  > out.reads ");
print "Peaks and reads intersected\n";

#### lets read the intersection and determine the status of the read
        my $statusc=0;

        open(INFILE,"<out.reads");
        while(<INFILE>){
    my @fields=split("\t",$_);
        my $id=$fields[8];
        if($readID_sig{$id} == 0 )
        { $statusc+=1; $readID_sig{$id}=1;}
        }
        close(INFILE);
        push( @statistics,
    join("\t",$n_sig,$limit_signal,$statusc,$n_mock,$limit_mock));
        print "Read status changes counted :",
    join("\t",$n_sig,$limit_signal,$statusc),"\n";

        #my $ask=getc; if($ask eq "y"){exit;}
        close(OUTFILE);
}
print join("\n",@statistics);

open(OUTFILE,">stats");
print OUTFILE join("\n",@statistics);
close(OUTFILE);

########################
sub fisher_yates_shuffle { my $array = shift; my $i;
        for ($i = @$array; --$i; ) { my $j = int rand ($i+1); next if $i == $j;
    @$array[$i,$j] = @$array[$j,$i]; } }
```

LISTING 6: Perl Script to Compute the Coverage of the Peaks by the Genomic Features

Appendix C. Supplementary Material for Chapter 4

Supplementary material related to the content of Chapter 4 is included in this Appendix.

C.25 Target Genes of BCL11B in the Mouse Brain

TABLE C.25: Genes Associated with BCL11B Peaks that Are Regulated by the Bcl11b in the Mouse Brain. Part 1

ACPL2	acid phosphatase-like
ADCY7	adenylate cyclase 7
ADORA2A	adenosine A2a receptor
AKNA	AT-hook transcription factor
ANXA6	annexin A6
APCDD1	adenomatosis polyposis coli down-regulated 1
APOBEC3D	apolipoprotein B mRNA editing enzyme catalytic polypeptide-like 3D
APPL2	adaptor protein phosphotyrosine interaction PH domain and leucine zipper containing 2
ARHGEF39	Rho guanine nucleotide exchange factor (GEF)
ARID5A	AT rich interactive domain 5A (MRF1-like)
ARRB2	arrestin beta 2
CAMTA1	calmodulin binding transcription activator 1
CARHSP1	calcium regulated heat stable protein 1 24kDa
CBFA2T3	core-binding factor runt domain alpha subunit 2 translocated to 3
CCBL2	cysteine conjugate-beta lyase 2
CD9	CD9 molecule
CFL1	cofilin 1 (non-muscle)
CKAP2L	cytoskeleton associated protein 2-like
COMMD3	COMM domain containing 3
CPT1A	carnitine palmitoyltransferase 1A (liver)
CST7	cystatin F (leukocystatin)
CXCL1	chemokine (C-X-C motif) ligand 1 (melanoma growth stimulating activity alpha)
CXCR5	chemokine (C-X-C motif) receptor 5
CYTH1	cytohesin 1
CYTH4	cytohesin 4
DUSP2	dual specificity phosphatase 2
DUSP22	dual specificity phosphatase 22
DUSP27	dual specificity phosphatase 27 (putative)
ELOVL5	ELOVL fatty acid elongase 5
EMID1	EMI domain containing 1
FKBP5	FK506 binding protein 5
GSK3B	glycogen synthase kinase 3 beta
HK1	hexokinase 1
HMGA1	high mobility group AT-hook 1
IDH3A	isocitrate dehydrogenase 3 (NAD+) alpha
IGHMBP2	immunoglobulin mu binding protein 2
IKZF2	IKAROS family zinc finger 2 (Helios)
IL12RB1	interleukin 12 receptor beta 1
IRF1	interferon regulatory factor 1
ITGA9	integrin alpha 9
KIF11	kinesin family member 11
KLF2	Kruppel-like factor 2 (lung)

Table C.25: Genes Associated with BCL11B Peaks that Are Regulated by the Bcl11b in the Mouse Brain. Part 2

LCP1	lymphocyte cytosolic protein 1 (L-plastin)
LDLRAP1	low density lipoprotein receptor adaptor protein 1
LRRC33	leucine rich repeat containing 33
LYL1	lymphoblastic leukemia derived sequence 1
MAPKAPK3	mitogen-activated protein kinase-activated protein kinase 3
MAPRE2	microtubule-associated protein RP/EB family member 2
MATK	megakaryocyte-associated tyrosine kinase
MCMBP	minichromosome maintenance complex binding protein
MDM1	Mdm1 nuclear protein homolog (mouse)
MSRA	methionine sulfoxide reductase A
NCAPG2	non-SMC condensin II complex subunit G2
NEIL1	nei endonuclease VIII-like 1 (E. coli)
PARP16	poly (ADP-ribose) polymerase family member 16
PCID2	PCI domain containing 2
PIK3CD	phosphatidylinositol-4 5-bisphosphate 3-kinase catalytic subunit delta
PLEKHF1	pleckstrin homology domain containing family F (with FYVE domain) member 1
PPM1M	protein phosphatase Mg2+/Mn2+ dependent 1M
PREX1	phosphatidylinositol-3 4 5-trisphosphate-dependent Rac exchange factor 1
PSMB10	proteasome (prosome macropain) subunit beta type 10
RASGRP1	RAS guanyl releasing protein 1 (calcium and DAG-regulated)
RUNX3	runt-related transcription factor 3
SEPHS2	selenophosphate synthetase 2
SH2B3	SH2B adaptor protein 3
SHISA5	shisa homolog 5 (Xenopus laevis)
SOCS1	suppressor of cytokine signaling 1
SSBP3	single stranded DNA binding protein 3
STK10	serine/threonine kinase 10
TASP1	taspase threonine aspartase 1
TBC1D10C	TBC1 domain family member 10C
TBC1D14	TBC1 domain family member 14
TBC1D16	TBC1 domain family member 16
TBC1D5	TBC1 domain family member 5
TCF7	transcription factor 7 (T-cell specific HMG-box)
TMBIM4	transmembrane BAX inhibitor motif containing 4
TNFRSF21	tumor necrosis factor receptor superfamily member 21
TPM4	tropomyosin 4
TTK	TTK protein kinase
USP20	ubiquitin specific peptidase
USP30	ubiquitin specific peptidase 30
VPREB1	pre-B lymphocyte 1
XRCC1	X-ray repair complementing defective repair in Chinese hamster cells 1
ZBTB7B	zinc finger and BTB domain containing 7B
ZFP36L2	ZFP36 ring finger protein-like 2
ZMIZ1	zinc finger MIZ-type containing 1

C.26 Genes associated with BCL11B Peaks

Table C.26: Genes Associated with BCL11B Peaks Containing the Motif 5'-GGCCGGAGG-3'

ACBD7	acyl-CoA binding domain containing 7
ADCY7	adenylate cyclase 7
ALG6	asparagine-linked glycosylation 6 alpha-1 3-glucosyltransferase homolog (S. cerevisiae)
CACNA1I	calcium channel voltage-dependent T type alpha 1I subunit
CACNA2D4	calcium channel voltage-dependent alpha 2/delta subunit 4
CBFA2T3	core-binding factor runt domain alpha subunit 2 translocated to 3
CHAMP1	chromosome alignment maintaining phosphoprotein 1
CHD7	chromodomain helicase DNA binding protein 7
CHI3L2	chitinase 3-like 2
DCAF17	DDB1 and CUL4 associated factor 17
DDIT4	DNA-damage-inducible transcript 4
DLC1	deleted in liver cancer 1
FLRT1	fibronectin leucine rich transmembrane protein 1
FXYD2	FXYD domain containing ion transport regulator 2
GET4	golgi to ER traffic protein 4 homolog (S. cerevisiae)
HDAC4	histone deacetylase 4
HSPH1	heat shock 105kDa/110kDa protein 1
IPO7	importin 7
KIF11	kinesin family member 11
KIFC1	kinesin family member C1
LSM4	LSM4 homolog U6 small nuclear RNA associated (S. cerevisiae)
LUC7L3	LUC7-like 3 (S. cerevisiae)
MLLT10	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog Drosophila) translocated to 10
MRPL38	mitochondrial ribosomal protein L38
NCL	nucleolin
NUP210	nucleoporin 210kDa
PIKFYVE	phosphoinositide kinase FYVE finger containing
PTPLA	protein tyrosine phosphatase-like (proline instead of catalytic arginine) member A
SH3RF3	SH3 domain containing ring finger 3
SRSF2	serine/arginine-rich splicing factor 2
TMC6	transmembrane channel-like 6
TRAPPC8	trafficking protein particle complex 8
TRIM41	tripartite motif containing 41
UBE2J1	ubiquitin-conjugating enzyme E2 J1
USP6NL	USP6 N-terminal like
ZBTB2	zinc finger and BTB domain containing 2
ZCCHC3	zinc finger CCHC domain containing 3

C.27 Example of Enrichment Analysis of BCL11B Binding Regions and Prioritization of the Associated Genes

In this section an example is presented of how genes - associated with the potential transcription factor binding sites - can be analyzed.

The biological contexts in which the BCL11B transcription factor might play a role can be explored by enrichment annotation analysis. The Genomic Regions Annotation Tool (GREAT) [McLean et al., 2010] performs annotation enrichment analysis for ChIP-Seq data using 20 annotation databases. GREAT associates the genomic regions with the genes by first assigning the regulatory region to the gene. The regulatory region is defined by the distances up and downstream of the TSS. The genes in which the regulatory region overlaps the genomic region are associated with that genomic region. GREAT performs queries in its annotation and pathway databases and identifies significantly enriched annotations of the associated genes.

GREAT was used to analyze the biological contexts of the genes associated with the BCL11B peaks by defining the gene regulatory region as ±50Kb up/downstream of the gene TSS. The enriched annotations in the databases of GO Biological Processes, the mouse MGI phenotypes, Human Disease ontology and Pathway Commons identified by GREAT are presented in Table C.27. The Pathway Commons ontology contains data on pathways from multiple sources including biochemical reactions, complex assembly, transport and catalysis events, and physical interactions involving proteins, DNA, RNA, small molecules and complexes. The enriched annotations in the GO Biological Processes category of the potential BCL11B targets in the Jurkat cells identified by GREAT program (Table C.27) align well with the known involvement of the BCL11B in regulation of the processes in the T-cell development.

Jurkat cells are a line of human T lymphocyte cells used to study acute T cell leukemia and T cell signalling. The enriched annotations in the mouse MGI phenotypes category of the BCL11B

Table C.27: Part1. Enriched Annotations Identified by GREAT of the Genes Associated to the BCL11B Peaks

Ontology	**Annotation Terms**
GO Biological Processes	immune response, hemostasis, posttranscriptional regulation of gene expression, **lymphocyte activation**, leukocyte activation, induction of apoptosis, T cell activation, induction of programmed cell death, leukocyte differentiation, lymphocyte differentiation, regulation of RNA stability, integrin-mediated signaling pathway, regulation of mRNA stability, T cell differentiation, negative regulation of caspase activity, positive regulation of T cell activation, regulation of T cell differentiation, response to laminar fluid shear stress, DNA damage response, signal transduction by p53 class mediator, regulation of phagocytosis
MGI phenotypes	abnormal immune system cell morphology, abnormal leukocyte morphology, abnormal bone marrow cell morphology/development, abnormal immune cell physiology, abnormal adaptive immunity, abnormal cell-mediated immunity, abnormal lymphocyte morphology, abnormal leukocyte physiology, abnormal T cell morphology, abnormal mononuclear cell morphology, abnormal lymphocyte physiology, abnormal leukopoiesis, abnormal myeloblast morphology/development, abnormal lymphocyte cell number, abnormal mononuclear cell differentiation, abnormal leukocyte cell number, abnormal lymphopoiesis, decreased hematopoietic cell number, decreased lymphocyte cell number, abnormal T cell number
Human Disease Ontology	acute myeloid leukemia, systemic lupus erythematosus, lung adenocarcinoma, Retroviridae infectious disease, herpes simplex, chronic myeloid leukemia, Human immunodeficiency virus infectious disease, thrombocytosis, **T-cell leukemia**, respiratory failure, peptic ulcer, lung small cell carcinoma, pulmonary edema, myelofibrosis, histiocytosis
Panther Pathways	T cell activation
Pathway Commons	TCR signaling in native CD8+ T cells, CXCR4-mediated signaling events, Hemostasis, TCR signaling in native CD4+ T cells, IL2-mediated signaling events, Platelet activation, signaling and aggregation, Thromboxane A2 receptor signaling, amb2 Integrin signaling, Hypoxic and oxygen homeostasis regulation of HIF-1-alpha, Semaphorin interactions, Adaptive Immune System, HIF-1-alpha transcription factor network, GPVI-mediated activation cascade, IL12-mediated signaling events, Validated transcriptional targets of AP1 family members Fra1 and Fra2, Aurora A signaling, Signaling by Aurora kinases, IL2 signaling events mediated by PI3K, E2F transcription factor network, Regulation of Telomerase

potential target genes comprise annotations related to abnormal immune system processes. The annotations in the human diseases category among other diseases contain the **T cell leukemia**. The enrichment analysis leads to the hypothesis that the genes associated with the BCL11B peaks may be relevant in the processes mediating leukemia.

Each enriched annotation is associated with genes by GREAT. Some genes within the group may be more important in the processes defined by the enriched annotations. To show the utility of the prioritization tool developed in this thesis the putative BCL11B target genes associated

with the **T cell leukemia** were prioritized with respect to **T cell leukemia** phenotype.

The gene candidates appearing within the Top 1% of the ranked whole genome with respect to the **T cell leukemia** phenotype are shown in Table C.27. Out of 101 associated genes in the list of the candidates 9 appear at the Top 1% of the ranked genome. Among the top ranking putative BCL11B target genes are NOTCH1, RUNX1, WT1. These genes are implicated in T ALL [Van Vlierberghe and Ferrando, 2012]. The NOTCH1 signaling represents prominent oncogenic pathway in T cell transformation in T-ALL [Paganin and Ferrando, 2011]. The mutations activating aberrant NOTCH1 signalling were identified in 60% of T-ALL patients. RUNX1 and WT1 are a tumor suppressor genes. Loss-of-function mutations were found in those genes in the T-ALL samples [Van Vlierberghe and Ferrando, 2012]. Loss-of-function (LOF) mutations and heterozygous deletions also are found in BCL11B in T-ALL suggesting that this may be an important pathogenetic event in T cell leukemogenesis [Van Vlierberghe and Ferrando, 2012]. In this analysis the results of the gene prioritization agree with the current knowledge in literature about molecular mechanisms underlying pathogenesis of T-ALL. It can only be speculated whether the tumor suppressor genes regulated by BCL11B are affected by the BCL11B LOF mutations in T-ALL.

TABLE C.27: Part 2. Ranking of the Potential BCL11B Target Genes with Respect to the **T Cell Leukemia** Phenotype

Gene	**Description**	**Gene Score**	**Gene Rank**
CTNNB1	catenin (cadherin-associated protein) beta 1 88kDa	0.0055	2
NOTCH1	notch 1	0.0388	9
IKZF1	IKAROS family zinc finger 1 (Ikaros)	0.0443	10
VEGFA	vascular endothelial growth factor A	0.1163	24
AKT1	v-akt murine thymoma viral oncogene homolog 1	0.1439	30
ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2 neuro/glioblastoma derived oncogene homolog (avian)	0.1550	33
HIF1A	hypoxia inducible factor 1 alpha subunit (basic helix-loop-helix transcription factor)	0.5481	121
TGFB1	transforming growth factor beta 1	0.6366	139
RUNX1	runt-related transcription factor 1	0.8082	178
STAT5A	signal transducer and activator of transcription 5A	1.0131	212
ESR1	estrogen receptor 1	1.0850	230
WT1	Wilms tumor 1	1.1293	238

Appendix D. Performance in the Prioritization of Disease Genes

A benchmark of 42 human genes known to be associated with human diseases is available for gene prioritization researchers through the supplementary material in the reference [Moreau and Tranchevent, 2012]. The phenotypes of these genes are defined by training genes and keywords. The data presented here is limited to only 26 cases as for some disease phenotypes an exact mapping of their defining keywords to MeSH disease terms was difficult to elicit. Thus, the developed prioritization algorithm was used to prioritize 26 disease genes whose phenotype keywords could be adequately matched to MeSH terms defining diseases. The ranking of these genes is shown in Table D.27.

The combined score is a rank ratio of genes within the whole ranked genome to the selected phenotype. In this setting, the whole genome can be thought of as a list of random genes against which the known disease gene is being ranked with respect to the selected disease phenotype. A summary of the combined scores of the genes is as follows: Minimum = 0.03, 1st Quartile = 1.94, Median = 10.13, Mean = 14.81, 3rd Quartile = 24.02, Maximum = 71.54.

The presented performance summary aligns well with the results achieved by Pinta, Endeavour and Suspects as reported in a study of the unbiased evaluation of gene prioritization tools (see Table 2 and Figure 1 in the reference [Bornigen et al., 2012]). The median rank ratio of

well-performing methods in that study is close to 11. For the data presented in Table D.27 the true positive rate within 5% as it is defined in reference [Bornigen et al., 2012] is 40%.

TABLE D.27: Rankings of the Disease Genes from Reference [Moreau and Tranchevent, 2012]

Phenotype	**Gene**	**Combined Gene Score (%)**	**Gene Rank**
Hernia, Diaphragmatic	HCCS	30	7401
Nasopharyngeal Neoplasms	TNFRSF	3.9	1226
Nasopharyngeal Neoplasms	MECOM	16.1	4404
Testicular Neoplasms	ATF7IP	24.02	6297
Testicular Neoplasms	DMRT1	10.13	2958
Crohn Disease	FUT2	40.02	9556
Crohn Disease	BCL3	0.75	244
Asthma	CSF1R	1.94	627
Craniosynostoses	GLI3	0.2	74
Cleft Lip	STOM	44.01	10608
Hypertelorism	UTRN	24.57	6750
Parkinson Disease	TRAF6	2.11	724
Liver Cirrhosis, Biliary	SPIB	26.93	7048
Liver Cirrhosis, Biliary	MMEL1	71.54	14695
Multiple Sclerosis	CRHR1	11.03	3196
Diabetes Mellitus, Type 2	UBE2E2	12.99	3646
Carcinoma, Squamous Cell	PLCE1	14.16	4241
Retinal Degeneration	SDCCAG8	17.3	5298
Alcoholism	HTR7	0.6	213
Psoriasis	TRAF3IP2	6.05	1814
Lung Neoplasms	TP63	1.41	484
Leprosy	IFNG	0.03	14
Microcephaly	WDR62	3.09	1103
Amino Acid Metabolism, Inborn Errors	CD320	0.42	154
Ehlers-Danlos Syndrome	CHST14	6.99	2267