



COPYRIGHT AND USE OF THIS THESIS

This thesis must be used in accordance with the provisions of the Copyright Act 1968.

Reproduction of material protected by copyright may be an infringement of copyright and copyright owners may be entitled to take legal action against persons who infringe their copyright.

Section 51 (2) of the Copyright Act permits an authorized officer of a university library or archives to provide a copy (by communication or otherwise) of an unpublished thesis kept in the library or archives, to a person who satisfies the authorized officer that he or she requires the reproduction for the purposes of research or study.

The Copyright Act grants the creator of a work a number of moral rights, specifically the right of attribution, the right against false attribution and the right of integrity.

You may infringe the author's moral rights if you:

- fail to acknowledge the author of this thesis if you quote sections from the work
- attribute this thesis to another author
- subject this thesis to derogatory treatment which may prejudice the author's reputation

For further information contact the University's Copyright Service.

sydney.edu.au/copyright

Development of novel software tools and methods for investigating the significance of overlapping transcription factor genomic interactions

by

Matloob Khushi

A thesis submitted in fulfillment of the requirements for the degree of

Doctor of Philosophy



THE UNIVERSITY OF
SYDNEY

Faculty of Medicine

The University of Sydney

© Matloob Khushi 2016

Declaration

I hereby declare that the work described herein this thesis submission is, to the best of the knowledge and belief, is original and is entirely my work, except where due acknowledgements have been made. The work was conducted, while I was pursuing a PhD degree program at the Faculty of Medicine, The University of Sydney at Westmead Millennium Institute, Westmead, NSW, under the supervision of Professor Christine Clarke and associate supervisor Dr. J. Dinny Graham. This thesis has not been submitted, wholly or in part, for the award of the higher degree to any other university or institution, and that all assistance/support received and all sources used in the completion of this thesis have been acknowledged.

Matloob Khushi

March 2016

Acknowledgement

I would like to express my deepest gratitude to my primary supervisor Professor Christine Clarke for providing me the opportunity to study under her supervision. I greatly appreciate and acknowledge that teaching biology to a student having an engineering and computer science background was very challenging. Therefore, I greatly thank and heartfelt appreciate her encouragement, assistance, counselling and support during this time. Under her supervision I learnt biology at a great length and gained vast knowledge and experience in conducting quality reproducible research. In addition, I believe this project would not have been possible without the enormous support of my co-supervisor Dr. J Dinny Graham. I greatly thank her for beautifully carving an information technologist into a bioinformatician and for her endless support in this regard. I am privileged to have learnt many skills from her. I also thank to Professor Christopher Liddle for his co-supervision and helpful feedback during the project. I greatly acknowledge and thank Dr Ashley Waardenberg who reviewed and proof-read my thesis.

I am grateful of my parents who provided me basic education and up bringing to get to this stage. I cannot thank enough my loving wife Iram Matloob for being a full-time mum and taking care of my four children during my studies. I thank my daughters Muntaha Matloob, Fizza Fatima, Imamah N. Matloob and son Huzefa M. Matloob for providing joyous company and encouragement during this time. I also thank all other family members and in-laws who encouraged and supported me all the time. I

thank my friends and colleagues including Associate Professor Jonathan Arthur, Dr Sharon Cunningham, Dr. Erdahl Teber, Dr. Mohammad Sohail Memon, Dr. Mufaz ullah, Dr. Harunor Rashid, Tram Doan, Jane Carpenter, Ahmad Al Odaib for their encouragement and time to time helpful feedback.

Dedication

This thesis is dedicated to my lovely wife Iram Matloob, who has always supported me in all walks of the life, and to my children Muntaha Matloob, Fizza Fatima, Huzefa Muhammad Matloob and Imamah Noor Matloob for their great company. I am truly thankful for having all of you in my life. I also dedicate this work to my father Khushi Muhammad and mother Mehmooda Sultana from whom I gained the first knowledge about this world.

Publications Arising From This Thesis

Journal Publications:

- **Khushi, M.** 2015. Benchmarking database performance for genomic data. 2015. J Cell Biochem, 116, 877-83.
- **Khushi, M.,** Clarke, CL. & Graham, JD. 2014. Bioinformatic analysis of cis-regulatory interactions between progesterone and estrogen receptors in breast cancer. PeerJ, 2, e654.
- **Khushi, M.,** Liddle, C., Clarke, CL. & Graham, JD. 2014. Binding sites analyser (BiSA): software for genomic binding sites archiving and overlap analysis. PLoS One, 9, e87301.

Oral Presentations at International Conferences:

- **Khushi M.,** Liddle C., Clarke CL., Graham JD. ‘Development of a genomic region database and analysis tool for the Galaxy platform.’ Beyond the Genome 2013, San Francisco, 3rd October 2013.
- **Khushi M.,** Clarke C., Graham JD. ‘Bioinformatic analysis of cis-regulatory interactions between progesterone and estrogen receptors in breast cancer’. International Conference on Bioinformatics (InCoB 2014) , Sydney, Australia 2nd August 2014

Oral Presentations at National Meetings:

- **Khushi M.,** Liddle C., Clarke CL., Graham JD. ‘Development of a genomic region database and analysis tool for Windows and the Galaxy’. Sydney Bioinformatics Research Symposium, 8 November 2013.
- **Khushi M.,** Liddle C., Clarke CL., Graham JD. Overlap analysis of genomic regions. Oral and poster presentation at Sydney Bioinformatics Research Symposium 9 Nov 2012, Garvan Institute, Sydney.

List of Figures

Figure 1-1: Enhancer regions.....	8
Figure 1-2: Chromatin has two broad structures.	12
Figure 1-3: In vitro techniques used to detect DNA–protein interactions.:	16
Figure 1-4: Strand-dependent bimodality in tag density.	21
Figure 1-5: ChIP-Seq computational analysis steps.	22
Figure 2-1: Various relative positions of the two regions.	52
Figure 2-2: Database structure employed to benchmark the performance of database systems.....	55
Figure 2-3: Comparison of region insertion performance.	58
Figure 2-4: Comparison of performance for identifying overlapping regions using RegMap and Geo functions.	59
Figure 2-5: Flow chart describing our decision for the selection of a database. Oval represents final successful choices.	63
Figure 2-6: Flow chart describing selection for a language for the development of BiSA.	68
Figure 2-7: BiSA Database Schema.	69
Figure 2-8: BiSA application architecture.	71
Figure 2-9: BiSA in a client-server architecture.....	71
Figure 2-10: Identifying three overlapping regions.	72
Figure 2-11: Restore database screen.	76
Figure 3-1: Import Datasets to Knowledge Base (KB).....	89
Figure 3-2: BiSA Select Datasets screen.	91
Figure 3-3: Analysis is the main overlap analysis tab of BiSA.	93
Figure 3-4: Venn diagrams in BiSA.	94
Figure 3-5: Statistical significance of overlapping regions.	96

Figure 3-6: Gene annotation.	97
Figure 3-7: New gene definitions.	98
Figure 3-8: Proximal features.	100
Figure 3-9: Administration of datasets.	101
Figure 3-10: BiSA for Galaxy web interface overview.	104
Figure 3-11: Uploading data into the Galaxy.	106
Figure 3-12: Clicking on the file name label reveals dataset information.	107
Figure 3-13: BiSA Import Datasets to KB (circled) automatically populates BED or GFF formatted datasets from the current history in the dropdown.	108
Figure 3-14: Submitting Import Datasets to KB form adds a new job in the queue shown in grey in the History.	109
Figure 3-15: Browse Datasets can be used to browse the dataset information available in the Knowledge Base (KB).	110
Figure 3-16: Analysis is the main screen where overlapping and non-overlapping datasets can be studied.	111
Figure 3-17: A section of Galaxy analysis screen.	112
Figure 3-18: HTML output of Galaxy analysis option.	114
Figure 3-19: Venn diagram is drawn as an HTML file and overlapping numbers used in the drawing are also shown.	115
Figure 3-20: Region sizes are calculated for a bin size of 100 and presented as a HTML file saved on the History.	116
Figure 3-21: The statistical summary, the overlap correlation value and the output file is displayed on the History.	117
Figure 3-22: Genomic features can be extracted with a given distance.	118
Figure 3-23: Example study of overlap between FOXA1 and FOXA3, CTCF and SA1, ZNF263 and c-Fos datasets.	121

Figure 3-24: Density plot showing distribution of distance from TSS of genes.	123
Figure 4-1: Distribution of datasets for four (mm8, mm9, hg18 and hg19) reference assemblies in the BiSA knowledge base.....	128
Figure 4-2: Peaks were called using MACS and HOMER tools for ER α datasets with E2, Tam and Fulv treatments.	133
Figure 4-3: Spread sheet for easy identification of degree of overlap among different datasets.....	135
Figure 4-4: Venn diagram showing degree of overlap among HNF4G, STAG1 and H3K4me1 datasets.	135
Figure 4-5: The degree of transcription factor overlap was transcribed into a network diagram.	137
Figure 4-6: Hierarchical clustering heat map showing correlation of 12 datasets in T47D cells using OCV calculated in Table 4.5.....	142
Figure 4-7: Transcription factor network in T47D breast cancer cell line.	143
Figure 5-1: Distribution of PR binding region sizes.....	154
Figure 5-2: Distribution of ER α binding region sizes.	154
Figure 5-3: Visualisation of ER α and PR overlapping common regions.	155
Figure 5-4: Example overlapping ER α -PR region.....	156
Figure 5-5: Statistical significance test using Genometricorr.....	159
Figure 5-6: Motif position distributions in ER α -PR overlapping regions.	161
Figure 5-7: Overlapping of 1,831 ERE and 8,259 PRE motif locations in 4358 common ER α -PR regions.	164
Figure 5-8: Comparison of PRE and ERE motifs with sequence logo generated from sequences of 285 common ERE-PRE motif locations.....	164
Figure 5-9: Venn diagram showing an overlap of H3K4me1, H3K4me3 and ER α -PR common regions.....	165

Figure 5-10: ER α -PR common regions-gene association.....	173
-------------------------------------------------------------------	-----

List of Tables

Table 1-1: Histone methylation marks identify the state of promoter of genes being active or silenced.....	14
Table 1-2: Publicly available ChIP-seq peak-calling software packages.....	27
Table 1-3: Software tools for motif analysis of ChIP-seq peaks and their uses.....	31
Table 1-4: Comparison of tools that operate on genomic regions.....	34
Table 1-5: List of tools that compute the statistical significance of overlapping regions.....	35
Table 2-1: Comparison of relational (SQL) databases and non- relational (NoSQL) databases.....	43
Table 2-2: Comparison of databases for operating system support, maintained by, license-type and maximum database size of seven top databases.....	47
Table 2-3: Overview of seven top computer languages.....	64
Table 4-1: OCV was calculated by selecting MACS datasets in first column (bold) as query while HOMER datasets were selected as reference.....	131
Table 4-2: OCV was calculated by selecting HOMER datasets in first column (bold) as query while MACS datasets were selected as reference.....	132
Table 4-3: Overlap Correlation Value (OCV) for PR datasets with various treatments.....	138
Table 4-4: OCVs among different ER α datasets.....	140
Table 4-5: OCVs among 12 compared datasets.....	141
Table 5-1: Motif analysis of PR regions.....	153
Table 5-2: Motif analysis of ER α regions.....	153
Table 5-3: BiSA Overlap Correlation Value (OCV) testing.....	158
Table 5-4: Known motif analysis of ER α and PR overlapping common regions.....	161

Table 5-5: De novo motif analysis of ER α and PR overlapping common regions.....	162
Table 5-6: Genes associated with ER α -PR common regions.	166
Table 5-7: Top 10 Gene Ontology (GO) biological process associated with ER α -PR common regions.....	168
Table 5-8: Top 10 Pathway Commons terms associated with ER α -PR common regions.	170
Table 5-9: Top 10 mouse phenotypes associated with ER α -PR common regions.	172
Table 5-10: Common differentially expressed transcripts by the treatment of estrogen or progesterone.....	174
Table 5-11: Top DAVID functional annotation of estrogen and progestin regulated transcripts that were associated with ER α -PR shared binding regions.....	175

List of Abbreviations

AR	Androgen Receptor
BiSA	Binding Sites Analyser (Software)
BPA	Bisphenol
ChIP	Chromatin Immunoprecipitation
ChIP-Seq	ChIP followed by high through put sequencing
E2	Estradiol
ER α	Estrogen Receptor Alpha
FDR	False Discovery Rate
FOXA1	Forkhead Box Protein A1
GEN	Genistein
GUI	Graphical User Interface
KB	Knowledge Base
OCV	Overlap Correlation Value
PR	Progesterone Receptor
RDBMS	Relational Database Management Systems
RegMap	Region Mapping (Algorithm)
RNA	Ribonucleic acid
RU486	Mifepristone
SQL	Structured Query Language
STAG1	Stromal Antigen 1
TES	Transcription End Sites
TF	Transcription Factor
TSS	Transcription Start Site

Table of Contents

Declaration.....	II
Acknowledgement.....	III
Dedication.....	V
Publications Arising From This Thesis.....	VI
List of Figures.....	VII
List of Tables.....	X
List of Abbreviations.....	XII
Abstract.....	XVII
Chapter 1: Introduction.....	1
1.1. Background.....	1
1.2. System Biology: from DNA to Organism.....	1
1.3. Gene Expression Regulation.....	2
1.4. Transcription Factors.....	3
1.5. Cis-regulatory Regions.....	6
1.5.1. Promoter Regions.....	6
1.5.2. Enhancer Regions.....	7
1.5.3. Silencer Regions.....	9
1.5.4. Insulator Regions.....	9
1.5.5. Locus Control Regions.....	10
1.5.6. Other DNA Regions.....	10
1.6. Chromatin.....	11
1.7. Epigenetics.....	12
1.8. Techniques Investigating Binding of Proteins to DNA.....	15
1.8.1. DNA Footprinting Assay.....	15
1.8.2. Electrophoretic Mobility Shift Assays (EMSA).....	17
1.8.3. Chromatin Immunoprecipitation (ChIP).....	18
1.8.4. ChIP-chip.....	19
1.8.5. ChIP-Seq.....	20
1.8.6. Peak Calling Algorithms for ChIP-Seq.....	24
1.8.7. Genomic Regions/Intervals.....	29
1.8.8. Analysis of Enriched Genomic Regions.....	29

1.9.	Challenges / Gaps.....	36
1.10.	Aims.....	37
Chapter 2: Selection of suitable computational resources to build a genomic database		
	39	
2.1.	Introduction	39
2.2.	Collection of Datasets.....	39
2.3.	Choice of Operating Systems	40
2.4.	Database Selection.....	41
2.5.	Relational or Non-Relational Databases	41
2.6.	Reviewing Relational (SQL) Databases.....	45
2.7.	Development of Overlapping Regions Algorithms.....	47
2.7.1.	Processing Genomic Regions.....	48
2.7.2.	Calculation of Base Pair Overlap (bp overlap)	49
2.7.3.	Calculation of Distance between Region Centres	52
2.7.4.	Extracting Overlapping Sections of Regions Common in Two Datasets ..	53
2.8.	Benchmarking Database Performance using RegMap	55
2.8.1.	Benchmarking for Insertion of Data	57
2.8.2.	Benchmarking for Identification of Overlapping Regions.....	58
2.8.3.	Searching and retrieving regions.....	60
2.8.4.	Advantages of RegMap over Geo functions	60
2.8.5.	Other Considerations for Choosing Between MySQL or PostgreSQL.....	62
2.9.	Language Selection for Writing BiSA	63
2.10.	BiSA Database Schema	67
2.11.	BiSA Application Architecture.....	70
2.12.	BiSA Charts	71
2.13.	Statistical Significance.....	73
2.14.	Gene Annotations	74
2.15.	BiSA on Sourceforge.net	74
2.16.	BiSA for Windows: Installation and Configuration Testing	74
2.17.	BiSA for Linux/Galaxy.....	77
2.18.	Tool Integration in Galaxy.....	78
2.19.	BiSA for Galaxy: Installation and Configuration Testing.....	80
2.20.	BiSA Developmental Issues	81
2.21.	Discussion.....	83

Chapter 3: Binding Site Analyser (BiSA): database resource, archival and tools to analyse genomic regions.....	87
3.1. Introduction	87
3.2. BiSA for Windows	88
3.2.1. Import Datasets to Knowledge Base (KB).....	88
3.2.2. Select Datasets	89
3.2.3. Analysis.....	90
3.2.4. Statistical Significance	95
3.2.5. Annotation.....	96
3.2.6. Proximal Features.....	98
3.2.7. Administration.....	99
3.3. BiSA for Other Platforms.....	101
3.4. BiSA for Galaxy: Web Interface Overview	102
3.4.1. Importing Datasets into BiSA	104
3.4.2. Browse Datasets	109
3.4.3. Analysis.....	110
3.4.4. Statistical Significance	116
3.4.5. Proximal Features.....	117
3.4.6. Annotation.....	118
3.5. BiSA Application Example	119
3.6. Discussion.....	124
Chapter 4: Significance of Transcription Factor Overlapping Regions.....	126
4.1. Introduction	126
4.2. Methods	127
4.3. Results	130
4.3.1. Validation of Overlap Correlation Value (OCV).....	130
4.3.2. Development of a Spreadsheet for Easy Identification of Degree of Overlap among Datasets	132
4.3.3. Directionality of Transcription Cooperation.....	135
4.3.4. Transcription Factor Networks in ER/PR Positive Breast Cancer.....	136
4.4. Discussion.....	145
Chapter 5: Bioinformatic analysis of cis-regulatory interactions between progesterone and estrogen receptors in breast cancer	148
5.1. Introduction	148

5.2. Methods	149
5.3. Results	152
5.3.1. Limited Overlap of ER α and PR Regions	155
5.3.2. Statistical Analysis of ER α -PR Overlap	156
5.3.3. Motif Analysis.....	159
5.3.4. ER α -PR Common Regions Interact on Enhancer Regions	163
5.3.5. Bioinformatics Enrichment Analysis of ER α -PR Common Regions	165
5.3.6. Differential Gene Expression Analysis by ER α /PR Binding.....	174
5.3.7. Gene Expression Regulation due ER α -PR Common Regions.....	174
5.4. Discussion.....	176
Chapter 6: Discussion	181
References.....	187
Appendix: Publications.....	233

Abstract

Identifying overlapping binding patterns of different factors is a major objective of genomic studies, but existing methods to archive large numbers of datasets in a personalised database lack sophistication and utility. In addition, there is no comprehensive database built-in algorithm at present to identify overlapping regions. Therefore I have developed a novel region-mapping (RegMap) SQL-based algorithm to perform genomic operations. Using RegMap I benchmarked the performance of PostgreSQL and MySQL databases. Benchmarking identified that PostgreSQL extracts overlapping regions much faster than MySQL. Insertion and data uploads in PostgreSQL were also better, although general searching capability of both databases was almost equivalent.

Using the RegMap algorithm I developed transcription factor DNA binding site analyser software (BiSA), for archiving of binding regions and easy identification of overlap with or proximity to other regions of interest. Results can be restricted by chromosome or base pair overlap between regions or maximum distance between binding peaks. BiSA is capable of reporting overlapping regions that share common base pairs; regions that are nearby; regions that are not overlapping; and average region sizes. BiSA can identify genes located near binding regions of interest and genomic features near a gene or locus of interest. BiSA can also calculate statistical significance of overlapping regions as an overlap correlation value. Overlapping results can be visualized as Venn diagrams. A major strength of BiSA is that it is supported by a comprehensive knowledge base of publicly available transcription factor binding sites and histone modifications, which can be directly compared to user data.

Using the BiSA knowledge base I identified that HNF4G nuclear receptor significantly collocate with cohesin subunit STAG1 (SA1) and H3K4me3 promoter marks in HepG2 cell line. In addition, I studied the overlap of various transcription factors and their binding sites in T47D cell-line and calculated statistical significance of the overlap using BiSA. This revealed that Progesterone Receptor (PR) binding as a result of RU486 (anti-progestin) treatment was significantly co-located with many other factors than PR binding as a result of progesterone treatment. It was also identified that (Estrogen Receptor Alpha) ER α and PR binding due to estrogen and progesterone treatment in T-47D cells share ~27% binding regions suggesting an interesting functional relationship between the receptors, which justified further study.

To investigate ER α and PR relationship further, we re-analysed raw data to remove any biases introduced by the use of distinct tools in the original publications. We identified 22,152 PR and 18,560 ER α binding sites (<5% false discovery rate) with 4,358 overlapping regions among the two datasets. BiSA statistical analysis revealed a non-significant overall overlap correlation between the two factors, suggesting that ER α and PR are not partner factors and do not require each other for binding to occur. However, Monte Carlo simulation by Binary Interval Search (BITS), Relevant Distance, Absolute Distance, Jaccard and Projection tests by Genometricorr revealed a statistically significant spatial correlation of binding regions on chromosome between the two factors. Motif analysis revealed that the shared binding regions were enriched with binding motifs for ER α , PR and a number of other transcription and pioneer factors. Some of these factors are known to co-locate with ER α and PR binding. In addition, gene expression analysis of ER α and PR revealed cell differentiation and apoptosis as top significant biological processes by the set of transcripts that were regulated by ER α -PR common regions. Therefore spatially close proximity of ER α binding sites with PR binding sites suggests that ER α and PR, in

general function independently at the molecular level, but that their activities converge on a specific subset of transcriptional targets.

In summary, the BiSA comprehensive knowledge base contains publicly available datasets describing transcription factor binding sites and epigenetic modification and provides an easy graphical interface to biologist for advance analysis options.

Chapter 1: Introduction

1.1. Background

The mission of biological sciences is to discover how the information encoded in cells drive the normal growth and developmental processes. Recent studies employing the latest experimental techniques have shown that various cellular factors work together in complex ways. A large range of genomic datasets from such studies have been deposited in public repositories. However, there is no easy way for biologists to compare their own results with previously published studies. Therefore, work presented in this thesis is an effort to develop a comprehensive genomic resource describing transcription factor binding sites and epigenetic modifications. The database resource hosts and integrates information from datasets in the public domain and combines it with in-house datasets. The resource and integrated tools provide novel ways of investigating and comparing uploaded datasets. The resource and tools are written for multiple computer platforms to cater for various research goals. An easy interface allows researchers to upload their datasets and provides novel options to analyse, statistically compare and annotate datasets.

As a background to the work, the following sections describe the roles of different DNA elements, DNA-binding proteins and other chromatin factors in the development and progression of diseases. I also describe the experimental techniques and other software tools that are being used in this domain.

1.2. System Biology: from DNA to Organism

Deoxyribonucleic acid (DNA) is the genetic material of all living organisms and carries the entire genetic information about an organism. Molecular messages transcribed from DNA encoded information are sent outside the nucleus to form proteins which regulate various cell functions. DNA transcription and its final translation into protein is a highly regulated process. Various genes are switched on or off at various stages of life and respond to

environment or drug stimulation. For example all human body parts contain the same genome, however, function of various body parts (eyes, heart, liver, stomach, brain etc) differ greatly. The answer lies in understanding how gene expression is regulated.

1.3. Gene Expression Regulation

Gene expression is the process by which a DNA code is synthesised into DNA products such as proteins and various ribonucleic acids (RNA). These products control various cellular functions and physiology. Changes in cell functions are mediated by regulation of gene expression on several levels. Disruption of gene expression patterns have been observed in development and progression of a variety of human diseases, including cancer, neurodegeneration and osteoporosis (Kim et al., 2013b; Jamieson et al., 2012; Golub et al., 1999; van't Veer et al., 2002). Comprehensive mapping of gene expression patterns by gene expression microarray has allowed the classification of disease sub-types, including cancers. (Golub et al., 1999; van't Veer et al., 2002). The expression of one gene can affect the expression of other genes, this all together gives a cell new biochemical or morphological properties (Beljanski, 2013). Up or down-regulation can cause serious diseases such as cancer and effect progression and relapse of the disease. For example, the over-expression of growth factor receptors frequently seen in cancer greatly impacts on chemotherapy response and relapses (Panasci et al., 2012).

Gene expression is regulated via several mechanisms which define which genes will be expressed and which genes will be silenced. One of these mechanisms is the binding of proteins to locus specific DNA that triggers synthesis of messenger ribonucleic acid (RNA). Messenger RNA is translated into various proteins that perform functions necessary for life. The other important mechanism that regulates gene expression is the regulation of RNA. RNA-level regulation is driven by interference with messenger RNA translation by small interfering RNA (siRNA) and micro RNA(miRNA) that results in

cytoplasmic localisation and degradation of mRNA (Hooper and Hilliker, 2013; Shabalina and Koonin, 2008). This inhibition of RNA molecules that results in post transcription gene silencing is called RNA interference (RNAi). RNAi plays a vital role in defense against viruses and transposable elements in eukaryotes (Shabalina and Koonin, 2008; Phillips, 2008).

DNA level regulation by the binding of various proteins, chromatin conditions and DNA modifications have been extensively studied with recent high throughput technologies. These studies have provided us many striking novel insights into regulation of cell functions at the DNA level and have identified that these regulatory processes are extremely complex. Our full understanding of these processes will help identify pathways that are important in regulating gene expression. Therefore the focus of this thesis is the analysis of genomic level transcriptional regulation by developing a novel tool.

1.4. Transcription Factors

It was established more than forty years ago that gene transcription is regulated inside the nucleus by binding of proteins to DNA (Galas and Schmitz, 1978). These proteins, known as transcription factors (TFs), are sequence specific DNA-binding proteins that regulate gene expression and function under the influence of epigenetic marks. TFS have been highly studied in developmental biology and are largely responsible for the development of body parts in animal morphology. Disruption of TF pathways lead to abnormalities in organisation and development. For example, genetic studies in the fruit fly (*Drosophila Melanogaster*), have established that absence of the Homeotic protein antennapedia transforms the antenna producing segment into a leg producing segment (Herke et al., 2005; Chen et al., 2013).

In addition to transcription factors' ability to bind DNA in a sequence-specific manner, they also interact with other factors, RNA polymerase, chromatin remodelling complexes

and small noncoding RNAs to initiate, enhance or repress transcription. Based on these characteristic factors can be classified into three main types: i) General transcription factors ii) activators, and iii) co-factors. General transcription factors such as Transcription initiation factors TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH form a pre-initiation complex on the promoter region and are required for almost all types of transcription (Claessens and Gewirth, 2004). Transcription is initiated by the binding of TFIID followed by binding of other factors and direction of RNA Polymerase II to the transcription start site (Claessens and Gewirth, 2004; Maston et al., 2006). *In vitro*, this assembly is sufficient to initiate a low level of (basal) transcription from DNA templates. However, *in vivo* transcription is enhanced by activating factors. Activators bind to specific DNA sequences and can stimulate transcription of inactive genes. Initially it was thought that transcription factor binding regions were located in sequences upstream of the core promoter (Ptashne and Gann, 1997), however, it is now accepted that transcription factor binding sites could be up/downstream of promoter or within the gene body (Taniguchi, 2014; Tsang et al., 2014; Stower, 2011). Co-factors can be part of various factor families such as homeodomain, Pit-Oct-Unc (POU), Pax, cysteine rich zinc finger, helix-loop-helix (HLH), basic leucine zipper (bZIP), forkhead, ETS (Pabo and Sauer, 1992; Latchman, 2008; Maston et al., 2006). They could form homodimers or heterodimers before or after binding and this can dictate the specificity to DNA binding sites (Claessens and Gewirth, 2004). Transcription factors recruit co-factors in a complex way to regulate transcription. For example, in prostate and breast cancer forkhead box protein A1 (FOXA1) facilitates the binding of androgen receptor (AR) and estrogen receptor alpha (ER α) respectively in regulating the transcription of AR or ER α dependent genes (Cheung and Kraus, 2010; Augello et al., 2011; Sahu et al., 2011; Fiorito et al., 2013). Another example of transcription factor cooperation is the interaction among Sox2, Oct4 and Nanog for

regulation of genes by binding at enhancer regions in embryonic stem (ES) cells (Chen et al., 2008; Zhang et al., 2011). Understanding how these factors work together will help identify novel target pathways used to regulate gene expression. Some factors contain two or more domains, for example, one could bind to DNA and another to other TFs to activate transcription like general control protein (GCN4) and glucocorticoid receptor (GR) (Murguia and Serrano, 2012; Meijnsing et al., 2009). Some factors also have transcription activation domains but lack DNA binding domains, e.g. herpes simplex virus VP16 protein contains an activation domain however cannot bind to DNA because it does not contain any DNA binding domain. Therefore it recruits **host cell factor (HCF) and the cellular factor Oct-1 DNA-binding domain** to activate transcription (Simmen et al., 1997; Goding and O'Hare, 1989). Similarly the Fos proteins (Fos, FosB, Fra-1 and Fra-2) alone cannot bind to DNA, and form heterodimers with the Jun proteins (Jun, JunB and JunD) to form a complex known as AP-1 transcription factor complex. The complex plays an important role in bone development, melanoma development and progression and in other important cell functions (Nakatsu et al., 2014; Zenz et al., 2008; Wagner, 2010; Kappelmann et al., 2014).

Treating diseases by regulating co-factors has shown promising results in some studies. For example, c-Myc is over-expressed in many human tumours which gives rise to numerous tumorigenic phenotypes (Wolf et al., 2015). Myc is activated by Max protein, therefore, to control Myc activity, reducing the availability of Max has been shown to be a promising target for cancer therapy (Berg, 2011). Therefore in recent years much of research has been done to identify transcription factor co-factors that work together in regulation of genes.

Transcription co-factor binding sites can be identified by various experimental techniques covered in the forthcoming sections, while, computational techniques have also been

employed to identify the co-activators, for example bioinformatic analysis of NKX2-1 binding sites identified the presence of AP-1, forkhead box (FOXA1) and estrogen receptor β (ESR2) motifs in human lung adenocarcinoma establishing that these factors work together in differential gene expression of LMO3 (Watanabe et al., 2013).

1.5. Cis-regulatory Regions

Transcription factors bind to specific sequences to regulate transcription, these regulatory regions are referred as cis-regulatory elements or cis-regulatory regions (Riethoven, 2010; Wray, 2007). Some cis-regulatory elements are conserved across many species (Wasserman et al., 2000; Bejerano et al., 2004). Some of these distant non-coding conserved cis-regulatory regions are shown to be acting as enhancers or silencers (Soccio et al., 2011; King et al., 2005; Shlyueva et al., 2014). These cis-regulatory elements and transcription factors controlling gene expression are fundamental gatekeepers of cell physiology. So understanding how interactions between TF and cis-regions occur and may be altered in disease is very important, however, there are lots of gaps in our knowledge in this area which need to be explored further.

In eukaryotes, protein-coding genes are regulated by a number of distinct transcriptional regulatory DNA elements, the most important of which are i) promoters ii) enhancers, iii) silencers and iv) insulators (Maston et al., 2006; Ogbourne and Antalis, 1998; Levine et al., 2014) described in the following sections.

1.5.1. Promoter Regions

Promoter regions are sequence specific templates that provision the transcription of genes by binding of RNA polymerase and other necessary transcriptional complex proteins. In eukaryote the promoters for RNA polymerase I and II are usually upstream of transcription start sites (TSS) but some promoters for RNA polymerase III lie downstream of the TSS, therefore, the promoter sequences define the direction of transcription (Cooper, 2000). In

recent studies DNA sequences from 500 to 3000 bp of TSS were considered as promoter regions (Meng and Vingron, 2014; He et al., 2014; Eckler et al., 2014). Sequences thousands of bases away from TSSs having elements that affect the transcription are called distal promoters (Riethoven, 2010; Delgado and Leon, 2006).

1.5.2. Enhancer Regions

Enhancers are DNA sequences either upstream or downstream of TSS which enhance or stimulate transcription by binding to specific proteins. Enhancers can be located on the same or on different chromosomes than the genes they target. Enhancers play an important role in differential gene expression by mediating transcription factor signals in a cell type-specific manner (Buecker and Wysocka, 2012) therefore they are also referred as cis-regulatory modules (CRM) (Shlyueva et al., 2014). The first enhancer region was discovered more than 35 years ago which was 72 bp SV40 DNA segment that increased the transcription of the β -globin gene 200 times in a transgenic assay (Banerji et al., 1981). Since then a large number of studies have characterised their properties; however, their role in various diseases has not been fully understood (Shlyueva et al., 2014; Spitz and Furlong, 2012; Calo and Wysocka, 2013; Wang et al., 2013).

An enhancer can span up to 500 base pairs or even much bigger in length and contains motif sequences to bind multiple TFs (Levine and Tjian, 2003). It is argued that distal enhancers recruit transcription factors forming a loop (Figure 1.1-A) bringing regulatory factors into close proximity and near to promoter regions so that the protein complex function combinatorial to activate transcription (Jiang and Levine, 1993; Palstra and Grosveld, 2012; Bulger and Groudine, 2011b). There is another theory (Figure .1.1-B) about enhancers working that the enhancer proteins actively scan along the chromatin fiber until it comes into contact with promoter complex (pink oval) and activates transcription. Such distal spatial interactions have been confirmed by studying various interactions

between enhancers and their target genes using chromosome conformation capture (3C) techniques and its variants circular chromosome conformation capture (4C), chromosome conformation capture carbon copy (5C) and Hi-C methods or by chromatin interaction analysis with paired-end tag sequencing (ChIA-PET, which is a combination of chromatin immunoprecipitation and various 3C-based methods). (de Wit and de Laat, 2012; Shlyueva et al., 2014). 3C is an experimental technique to study spatial organisation of long genomic regions in living cells (Gavrilov et al., 2009).

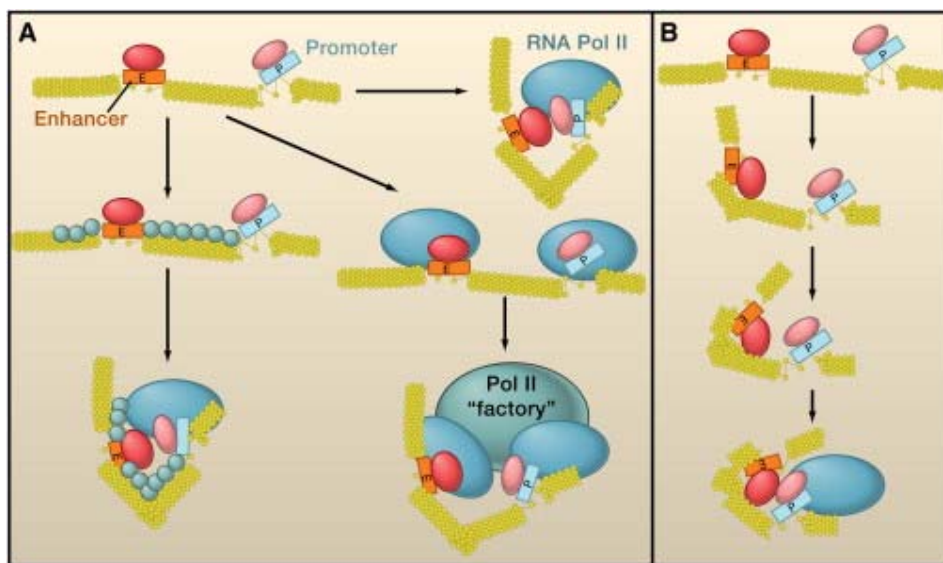


Figure 1-1: Enhancer regions. Enhancers shown as rectangle E recruit transcription factors forming a loop bringing regulatory factors into close proximity and near to promoter regions (shown as rectangle P). The protein complex recruit RNA Pol II to activate transcription. B) The enhancer binds to protein complex (red oval) which scans along the DNA in search of the promoter complex (pink oval) and activates transcription. Modified from (Bulger and Groudine, 2011b).

It has been shown that interruption of mammalian enhancer function greatly affects the development and progression of diseases (Ong and Corces, 2012). For example, mutations and insertions in long-range enhancer sequences regulating expression of the sonic hedgehog regulator ZRS, develop several forms of preaxial polydactyly in humans, mice,

cats and chickens (Albuisson et al., 2011) demonstrating the importance of those sequences in ensuring normal developmental regulation. The locations of enhancer elements are varied for different genes that they regulate. They can be near promoter regions, within the introns of the regulated genes, in the body of neighbouring genes or even on a different chromosome.

Enhancers regions can be predicted by specific histone modifications identified by chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) (Maston et al., 2012) (more detail can be found in the Epigenetics and ChIP-seq section).

1.5.3. Silencer Regions

Silencer regions suppress gene expression by the binding of repressor proteins. Silencers can be up or downstream of a TSS, within introns or exons (Ogbourne and Antalis, 1998). Silencers work exactly opposite to enhancers by turning off active genes (Maston et al., 2006). Like enhancers, silencers often act at a distance that can reach 100 kb (kilo bases) to repress promoter activity. For example, polycomb (PcG) proteins bind to silencer DNA sequences inhibiting the expression of Hox genes through early development in *Drosophila melanogaster* (Dean, 2011; Kyrchanova and Georgiev, 2014). Some DNA regions can act both as enhancer or silencer regions depending on what proteins are bound to them. For example consensus sequence 5'-CACGTG-3' which is known as E box when it is bound by MYC/MAX complex, transactivates its target genes implicated in the crucial cellular processes such as cell cycle regulation, proliferation, metabolism and mitochondrial biogenesis. However, when E box is bound to Mad/Max dimer it suppresses transcription (Dang, 2012; Taniguchi et al., 2014).

1.5.4. Insulator Regions

Enhancer and silencer regions can act on a number of genes at long distances, however, their activity is blocked by the binding of proteins on specific DNA regions known as

insulators. Chromatin forms loops which divides chromosomes into biological domains separated by insulator regions. These insulator regions act by binding to insulator proteins such as transcriptional repressor CTCF (CCCTC-binding factor) (Fiorentino and Giordano, 2012).

Insulators play a key role in various cell physiologies such as embryonic, neuronal and haematopoietic differentiation. In flies a number of insulator proteins are known however in vertebrates the CCCTC-binding protein CTCF is the only known insulator protein to date. CTCF prevents undesirable interaction between active and inactive genomic regions by binding to insulator sequences, and it can also protect particular genes from enhancer activity (Herold et al., 2012; Dean, 2011).

1.5.5. Locus Control Regions

“A Dictionary of Biomedicine” defines a Locus control region (LCR) as a non-transcribed region that contains the promoters and enhancers which regulate the expression of a particular gene (Lackie, 2010). The LCR was first described in transgenic mouse studies where it was identified that beta-globin LCR regulate expression of several genes (Palstra et al., 2008; Gerstein et al., 2007).

LCR are also considered long-range cis-acting sequences that effect gene regulations. LCR influence dynamic intra- and interchromosomal interactions between specific genetic loci that regulate transcriptional initiation or silencing of these loci (Spilianakis et al., 2005).

1.5.6. Other DNA Regions

There are other genomic regions which can influence gene transcription such as exons, introns, 3'-UTRs and intergenic regions. Historically the non-coding genome was believed to be ‘junk’, however, recent studies including the human genome project have revealed many previously undescribed genes, and new roles for non-coding DNA are constantly emerging (Shen et al., 2013; Djebali et al., 2012).

1.6. Chromatin

In mammalian cells, DNA is packaged into a very compact arrangement, termed chromatin, that stores about 2 metres of DNA into a nucleus of approximately 6 micrometres diameter (Alberts, 2008; Le Guezennec et al., 2005). Chromatin is the combination of DNA and other proteins in the nucleus, where DNA is wrapped twice around octamer histone molecules (pair of each H2A, H2B and H3) called nucleosomes (Latchman, 2008). The dense nature of chromatin regulates the binding of regulatory proteins and RNA polymerase, affecting the silencing or expression of genes as a result. The binding of proteins to specific DNA sequences depends on the availability of target regions that are not tightly compacted (euchromatin) and the way the DNA is packaged in the chromatin structure (Figure 1.2). The condensed form of chromatin known as heterochromatin makes the DNA inaccessible for most protein binding preventing transcription in those regions. Therefore chromatin must remodel and must be in a euchromatin state in order for gene expression to take place (Phillips and Shaw, 2008; Russ et al., 2012).

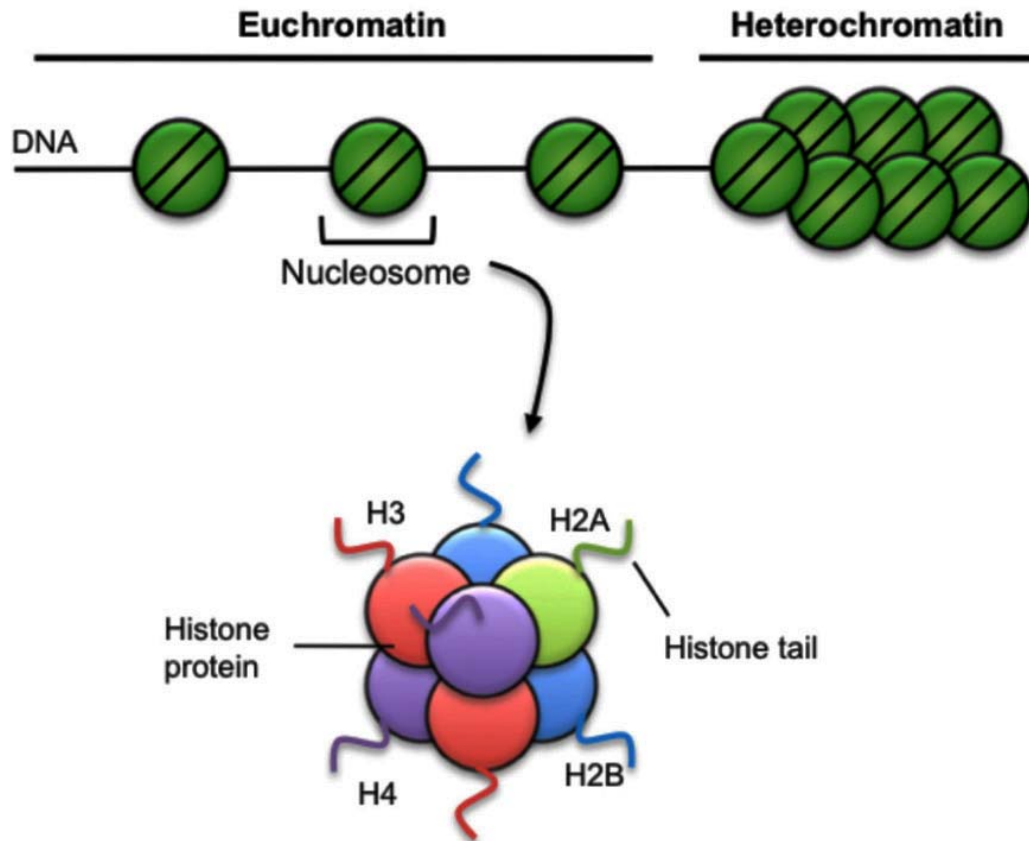


Figure 1-2: Chromatin has two broad structures. The first is euchromatin, characterized by sparse nucleosome density and is generally associated with active gene transcriptional activity. Heterochromatin is characterized by high nucleosome density, is very compacted and is generally associated with repression of gene transcription. Nucleosomes consist of 147 bp of DNA wound 1.65 turns around a complex of histone proteins, comprising two each of the H2A, H2B, H3, and H4 histone variants. Each histone has a soluble amino terminal tail that can be covalently modified by specific epigenetic marks discussed in the forthcoming section. Modified from (Russ et al., 2012).

1.7. Epigenetics

In addition to DNA sequence, transcription factor binding is profoundly influenced by the cell-specific epigenome, the pattern of post-translational modifications to histones and

other chromatin proteins, and chemical modifications to DNA, which ultimately direct nuclear chromatin structure and accessibility to transcribed regulation (Holliday, 2006; Carey and Smale, 2001). There are two main types of epigenetic modifications i) DNA methylation and ii) histone modifications. In DNA methylation a methyl group is added to cytosine nucleotide. The specific cytosine is always located next to a guanine nucleotide that is linked by a phosphate; this is called a CpG site. In histone modifications, histone can be modified by either methylation or acetylation. These modifications dictate whether the chromatin state is active (not condensed) or not active (heterochromatin) (Simmons, 2008; Egger et al., 2004; Jones and Baylin, 2002).

In cancer, increased DNA methylation (hypermethylation) of proximal promoter regions is associated with inappropriate transcriptional silencing of genes and is found in virtually every type of human neoplasm (Jones and Baylin, 2002; Baylin and Herman, 2000; Jones and Laird, 1999). Promoter hypermethylation has also been found in tumour-suppressor genes such as the BRCA1 gene which is silenced by promoter hypermethylation in primary breast and ovarian carcinomas (Esteller et al., 2000). Other example of hypermethylation is the inactivation of tumour suppressor gene ppENK in development and progression of pancreatic carcinogenesis (Jones and Baylin, 2002; Yang et al., 2013).

Histone H3 is the most characterised histone protein and majority of these modifications exist in the N-terminal tail. Trimethylation of Histone H3 lysine K4 (H3K4me3) and H3K79me3 and H3K27me1, H3K9me1 and H4K20me1 are associated with gene activation, however, H3K27me2 and H3K27me3 are silencing marks (Table 1.1) (Wei et al., 2009). Some genomic regions co-localise activation (H3K4me3) and silencing marks (H3K27me3), known as bivalent chromatin marks (Azura et al., 2006). These bivalent modifications could silence developmental genes in embryonic stem cells while keeping them poised for activation. (Roh et al., 2006; Bernstein et al., 2006). Its has been shown

that epigenetic marks H3K4me1,-2, -3; H3K9me1; H3K36me3; and H3K27me1 or –ac exhibit specific characteristics of enhancers (Maston et al., 2012; Creyghton et al., 2010; Pekowska et al., 2011; Zentner et al., 2011)

Activation Marks	H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27me1, H3K9me1
Silencing Marks	H3K9me2, H3K9me3, H3K27me2 H3K27me3,

Table 1-1: Histone methylation marks identify the state of promoter of genes being active or silenced.

Epigenetic modifications are inheritable and therefore are considered to drive many complex cell functions (Simmons, 2008; Egger et al., 2004). The tightly wound packaging of DNA around nucleosomes creates a barrier for reading and interpreting the stored DNA-sequence information (Le Guezennec et al., 2005). Access to this information is controlled by covalent post-translational chromatin modifications such as acetylation, methylation, phosphorylation. Therefore these alterations play a key role in transcription regulation (Bauer et al., 2002). A number of enzymes have been identified which catalyse the addition and removal of these modifications to histone proteins (Collas and Dahl, 2008). Different combinations of modification marks establish whether nucleosomes will be remodelled to activate or de-activate gene expression. DNA transcription occurs only if the chromatin is not tightly compacted (Figure 1.2) and DNA is available to bind transcription factors. A classic example in female mammals, only one of the two X chromosomes is transcriptionally active to compensate for the difference in dosage of X-linked genes between males and females (Goto et al., 2002; Egger et al., 2004). This is achieved by

epigenetic modifications that silence one of the two X chromosomes. Such deactivation of gene expression by epigenetic marks is called epigenetic silencing.

Recent studies have examined the effects of epigenetic drug treatments (Azad et al., 2013; Consalvi et al., 2011). Though a great deal of success has not been achieved yet, experimental and preclinical evidence results are promising. One of the barriers to using epigenetic drugs is their multiple effects on various pathways, however this could be a favourable property of this treatment, as tumour cells also exhibit abnormal regulation of many diverse pathways (Azad et al., 2013; Lawrence et al., 2015).

1.8. Techniques Investigating Binding of Proteins to DNA

Apart from gene regulation, DNA binding proteins play a key role in the regulation of DNA replication and recombination, repair, segregation, chromosomal stability, cell cycle progression, and epigenetic silencing (Das et al., 2004). Therefore various experimental techniques have been developed to investigate transcription factor binding to DNA. The following sections describe the most used experimental methods that tremendously increased our understanding of DNA-protein interactions.

1.8.1. DNA Footprinting Assay

Based on electrophoresis, Galas and Schmitz developed DNA footprinting in 1978 to study sequence-specific binding of proteins to DNA (Galas and Schmitz, 1978; Brenowitz et al., 1986). In this assay, the specific DNA sequence, whose protein binding properties is being studied, is radioactively or fluorescently labelled (Hampshire et al., 2007). DNA fragments are mixed with the protein under study. Sufficient time is given to form DNA-protein complexes. The complexes are then electrophoresed on a denaturing polyacrylamide gel, which separates the resulting DNA fragments according to their size. After electrophoresis the position of the DNA fragments are visualised by autoradiography (Leblanc and Moss, 2001). Figure 1.3 courtesy of Barski et al. (Barski and Zhao, 2009) explains the technique.

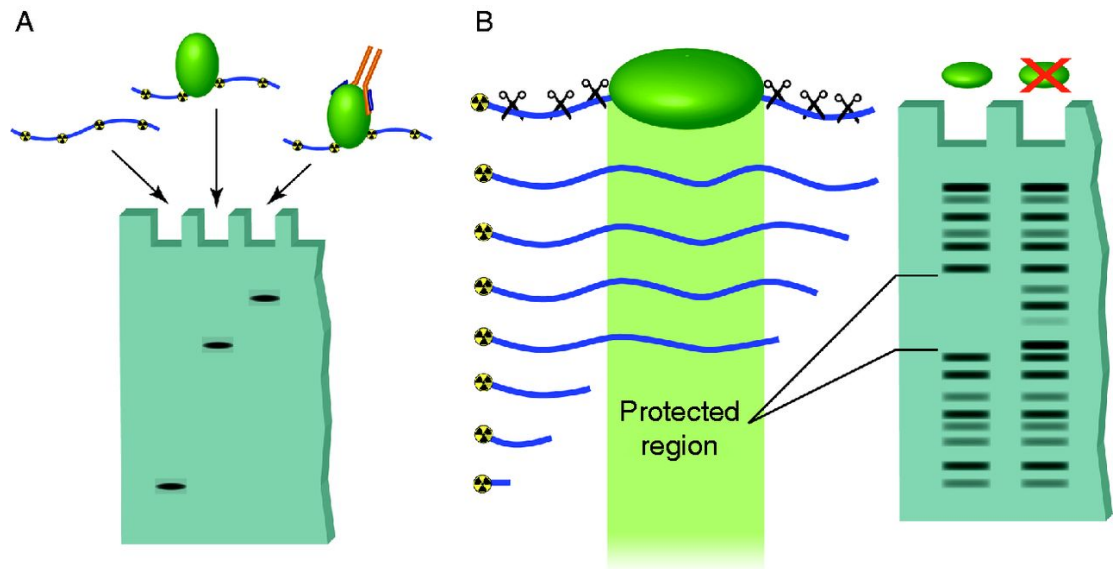


Figure 1-3: In vitro techniques used to detect DNA–protein interactions.: (A)

Electrophoretic mobility shift assays can be used to determine direct binding between a specific sequence of radioactively labelled DNA and a purified protein. Unbound DNA, termed free probe, migrates at a relatively low molecular weight in the agarose gel.

Binding of protein to this sequence results in the DNA band shifting to a high molecular weight region. Addition of an antibody that recognizes the bound protein causes an even greater shift in mobility, called supershifting. This assay can also be used with protein complexes to detect indirect protein–DNA interactions.

(B) DNase footprinting assays allow identification of regions of DNA bound by proteins. A DNA oligomer is radioactively labeled on one end and mixed with the protein of interest. The DNA is then digested by a DNA endonuclease (DNase). The regions of DNA that are bound by proteins are protected from digestion. When the DNA is run out on a gel, the protected region shows up as a break in the laddering produced by DNase digestion (Vinckevicius and Chakravarti, 2012a).

Once the sequence of a binding motif is identified from DNA footprinting then computational approaches can be used to identify genome-wide binding sites of transcription factors (Wasserman and Sandelin, 2004). For example, Laurell et al. used computational transcription factor-binding site predictions to suggest the sonic hedgehog ZRS limb enhancer mutation that creates new binding sites and causes ectopic gene expression (Laurell et al., 2012). However due to the repetitive nature of DNA sequence and short length of consensus sequences or motifs, typically from 6-15 base pairs, their use in genome-wide computational detection returns a large number of false positive regions. Most transcription factor binding sites identified by computational approaches fail to represent true *in vivo* binding sites, yet low significance variant motifs can often bind transcription factors *in vivo* (Barski and Zhao, 2009). This is likely to be due to the fact that many binding sites are highly cell context specific and their availability for binding is influenced by other factors besides the presence of a consensus motif sequence, such as chromatin structure and accessibility, or the expression of transcriptional cofactors. Today genome-wide protein-protein and DNA-protein associations are more widely studied by chromatin immunoprecipitation (ChIP) based assays.

1.8.2. Electrophoretic Mobility Shift Assays (EMSA)

Electrophoretic mobility shift assays (EMSA), or gel shift, is an *in vitro* experimental technique to study DNA and protein interactions. Gel electrophoresis is also an important component in this method. In the experiment the shift of the negative control lane that has only radio labelled genetic material and a lane with mixture of DNA-protein complex is compared. If protein binds to DNA the complex migrates slowly and a shift can be seen when the gel is dried and placed against X-ray film.

1.8.3. Chromatin Immunoprecipitation (ChIP)

ChIP was developed by Varshavsky and colleagues in 1988 to study protein-DNA interactions (Solomon et al., 1988). ChIP technique provides accurate information about the binding of TF, cofactor recruitment and epigenetic status during activation or repression of a DNA sequence and associated regulatory regions. This is an *in vivo* technique in which protein of interest is allowed to interact with chromatin in a living cell or tissue. Chromatin is fragmented and then immunoprecipitated (IP) using a highly specific antibody against the protein of interest (Mukhopadhyay et al., 2008).

There are two general types of ChIP procedures, Native and Cross-Linked, based on whether DNA-protein binding is cross-linked with formaldehyde (Das et al., 2004) or not. Generally, in Native-ChIP (NChIP) proteins are unfixed and are fragmented by micrococcal nuclease digestion whereas in Cross-Linked ChIP (XChIP) proteins are cross-linked with formaldehyde and fragmented by sonication (Le Guezennec et al., 2005). Finally the DNA is reverse-cross-linked, purified and enrichment of ChIP-ed DNA is analysed by quantitative real-time polymerase chain reaction (ChIP-qPCR), microarray (ChIP-chip) or next generation sequencing (ChIP-Seq).

There are a number of advantages and disadvantages for both of the techniques. In NChIP, antibody specificity is predictable and immunoprecipitation is very efficient as the antibody is able to bind effectively to the target antigen, therefore, precipitated DNA can be studied without further PCR amplification (Das et al., 2004; O'Neill and Turner, 2003). However, on the other hand, generally NChIP is limited to histone and histone modifications studies, as non-histone proteins are generally less tightly bound to DNA and may disassociate during sample preparation. Secondly nuclear digestion by micrococcal nuclease favours some genomic sequences over others, resulting in un-equal detection of those favoured genomic regions and thirdly as the nucleosomes are not fixed their

rearrangement can also occur. Overall NChIP requires more care than XChIP as the interactions are not fixed, however some studies have preferred this method for small cell numbers and higher sensitivity (O'Neill and Turner, 2003; Gilfillan et al., 2012).

XChIP is one of the most useful techniques for studying *in vivo* gene regulation by formaldehyde cross-linking of proteins to proteins and proteins to DNA, followed by immunoprecipitation of the fixed material (Orlando, 2000). XChIP is suitable for non-histone proteins and chromatin associated factors that bind weakly or indirectly to the DNA as cross-linking will fix these interactions. Cross-linking minimises nucleosome rearrangements as interactions are stabilised and there is less variability between experiments (Orlando et al., 1997). However limitations include the dependence on the availability of a highly specific antibody. Inefficient antibody binding may be observed due to the epitopes that the antibody need to recognize in the XChIP may be disrupted or destroyed by formaldehyde cross-linking and it may be necessary to test a variety of different antibodies to choose the best one (Mukhopadhyay et al., 2008). DNA sizes can vary widely due to over-fixation by formaldehyde. Another disadvantage of XChIP is that weak transient protein interactions can be fixed which can lead to false positive with results (Orlando et al., 1997).

1.8.4. ChIP-chip

ChIP-chip or (ChIP-on-chip) refers to the combination of ChIP assay with DNA microarray technology (ChIP-chip) to facilitate large-scale or genome wide analysis of the location of DNA-bound proteins (Ren et al., 2000). An oligonucleotide microarray chip contains thousands of unique DNA sequences, which act as probes for DNA which is purified and labelled after the ChIP assay. The immunoprecipitated enriched DNA is amplified and labelled with a fluorescent dye (Cy5) or biotin and a sample of DNA that is not enriched by immunoprecipitation labelled by a different fluorophore (Cy3). In two-

colour microarrays, both samples are hybridized to the microarray. The hybridized chip is placed in a laser scanner that activates the fluorescent dye present in the samples showing red and green spots (Ren et al., 2000). Various computational and statistical approaches are designed to capture and calculate the red and green ratio to normalise the enriched and reference channels (Royce et al., 2005; Lee et al., 2006; Elnitski et al., 2006). This allows the identification of DNA sequences that are over-represented in the ChIP sample, representing predicted binding sites.

The ChIP-chip technique has successfully been applied to map the genomic location of many transcription factors such as progesterone receptor (PR) (Tang et al., 2011), estrogen receptor alpha (ER α) (Hurtado et al., 2008), androgen receptor (AR) (Yu et al., 2010), forkhead proteins (FoxA1) (Lupien et al., 2008), GATA binding protein 3 (GATA3) (Hua et al., 2009), and histone modifications and has defined many new biological insights.

1.8.5. ChIP-Seq

ChIP followed by high throughput massively parallel sequencing (ChIP-Seq) has unveiled locations of protein binding sites and epigenetic marks genome-wide (Park, 2009). The ChIP-seq approach produces tens to hundreds of millions of short sequence reads usually referred to as 'tags' (Figure 1.4). In addition to the sequence itself, most platforms assign a quality score to each base, which is proportional to the estimated probability of an incorrect base call at that position (Cock et al., 2010). Low quality reads can be removed or trimmed. The remaining reads are then aligned with a reference, usually the genomic sequence of the organism used to generate the original ChIP samples. This is a computationally intensive step, and specific and highly efficient software tools have been developed to enable the task.

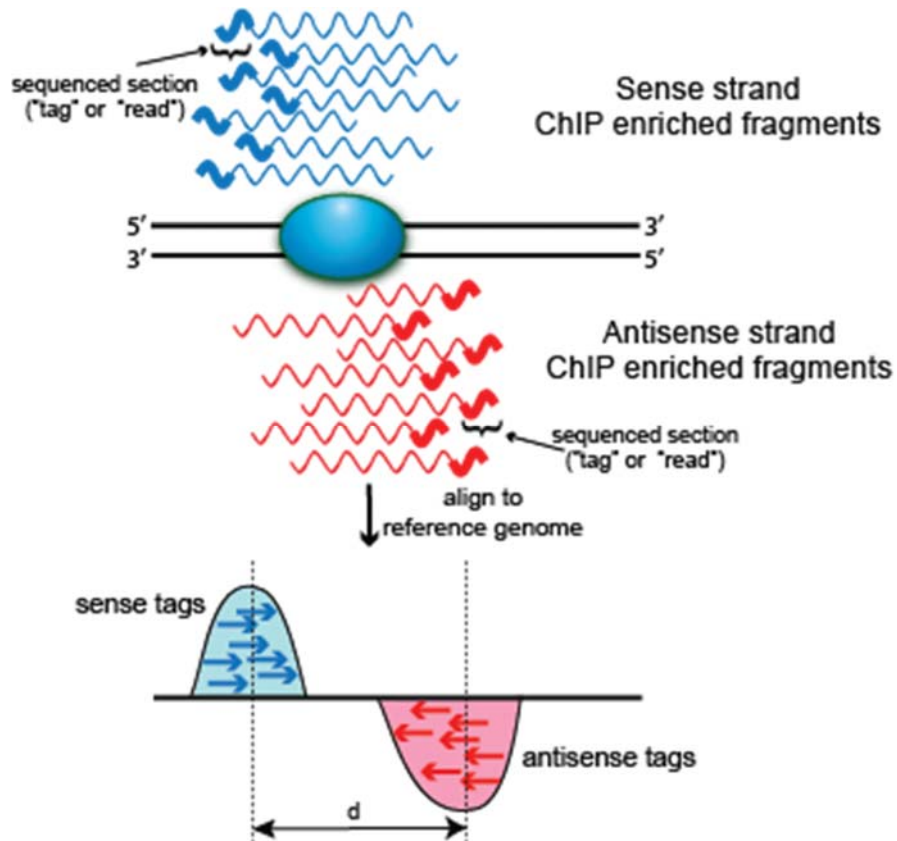


Figure 1-4: Strand-dependent bimodality in tag density. The shaded blue oval represents the protein of interest bound to DNA (solid black lines). Wavy lines represent either sense (blue) or antisense (red) DNA fragments from ChIP enrichment. The thicker portion of the line indicates regions sequenced by short read sequencing technologies. Sequenced tags are aligned to a reference genome and projected onto a chromosomal coordinate (red and blue arrows). Sequence-specific binding events (e.g. transcription factors) are characterized by “punctuate enrichment” (Pepke et al., 2009) and defined strand-dependent bimodality, where the separation between peaks (d) corresponds to the average sequenced fragment length. Inspired by Jothi et al. (Jothi et al., 2008) modified from Wilbanks and Facciotti (Wilbanks and Facciotti, 2010).

Bowtie (Langmead et al., 2009) and BWA (Li and Durbin, 2009) are two widely used software programmes for sequence alignment. The choice of program and the alignment parameters used affect the number of mapped reads. Specialized “peak calling” software is used to collate the number of sequence reads detected at each base position in the alignment reference. Tags are counted at each location and significant peaks, representing higher numbers of aligned tags at that location relative to a background level (Figure 1.4) are identified by peak-calling software. There are a number of software tools available for peak-calling (Kim et al., 2011; Pepke et al., 2009). The numbers of peaks reported by these programs are highly dependent on the parameters used (Kim et al., 2011; Pepke et al., 2009). Peak calling algorithms are discussed in the forthcoming section. Figure 1.5 diagrammatically explains ChIP-Seq computational analysis steps.

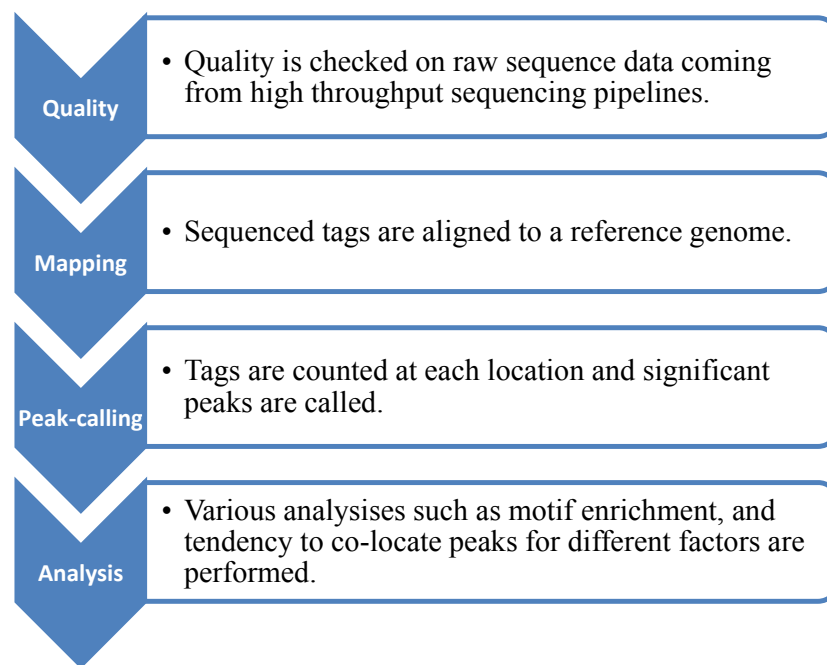


Figure 1-5: ChIP-Seq computational analysis steps. Output of one step becomes the input of other step. Sequenced reads (tags) from the sequencing platform the quality of reads are ensured, reads are mapped to a reference genome followed by peak-calling and finally the peaks are analysed.

Transcription factor peaks usually span few hundred base pairs such as Estrogen Receptor alpha (ER-alpha) while histone modification (e.g. H3K4me1) peaks are usually spread up to many kilobase pairs, however, in some cases, such as RNA polymerase II a mix of small and long regions is detected (Pepke et al., 2009).

ChIP-Seq has many advantages over ChIP-chip technology by providing a better signal-to-noise ratio (Johnson et al., 2008; Ho et al., 2011), higher specificity, sensitivity and comprehensive coverage of transcription factor binding sites or epigenetic markers across the genome (Kim et al., 2011). ChIP-seq also generally identifies a larger number of more narrowly focused binding intervals (often referred to as peaks) compared to ChIP-chip.

A disadvantage of ChIP-Seq is that a great deal of computational work is required to analyse ChIP-Seq data. Box 1.1 lists advantages of ChIP-Seq technique over ChIP-chip technology.

ChIP-Seq versus ChIP-chip

Advantages over ChIP-chip

- Higher resolution
- Less noise, no cross hybridisation, higher dynamic range
- Greater genome-wide coverage
- Cheaper
- Small amount of sample is required
- Less amplification is required

Disadvantages

- Big datasets are generated, therefore need large storage space
- Data-analysis is challenging, highly sophisticated bioinformatics tools and high computation power are required

Box 1.1: Comparison of ChIP-Seq and ChIP-chip technologies.

Generally ChIP-Seq is a superior technology and has been employed by most recent studies on transcription factor binding sites and histone modifications (Furey, 2012; Vinckeivicius and Chakravarti, 2012b).

Both techniques are being employed for genome-wide profiling of various factors which has provided insight into their role in health and diseases. Therefore these assays have played fundamental role in unveiling striking novel biological findings which was not possible with other platforms (Vinckeivicius and Chakravarti, 2012b; Carlberg, 2014).

1.8.6. Peak Calling Algorithms for ChIP-Seq

After aligning the read tags, the major step is to find genomic regions that have a significantly higher number of tags than the background. Peak calling identifies the enrichment regions where the protein of interest binds on DNA. There are numerous open source peak-calling programs available, however the challenge of selecting a suitable tool for a study remains confronting (Laajala et al., 2009; Pepke et al., 2009; Wilbanks and Facciotti, 2010). Most early peak-calling software identified peaks by merely counting the regions of the genome having high read density in the ChIP sample (Johnson et al., 2007; Robertson et al., 2007). This method provides general peak identification, however the identification of the DNA binding location is not exact. Later software have taken the shape of the peak, directionality of sequencing reads and statistical significance of the peak compared to background into account (Valouev et al., 2008; Goecks et al., 2010a). For example FindPeaks software provides options to refine peaks by trimming, identify directional reads ignoring fragments after or before the peak on the forward or reverse strand, sub-peak identification and separating multiple peaks joined together (Fejes et al., 2008). SiSSRs (Site Identification from Short Sequence Reads), MACS (Feng et al., 2012) and QuEST (Valouev et al., 2008) software shift tags by half of the estimated fragment size towards the centre of the peak to result in sharper peaks, while tag shifting is an option

in ERANGE and FindPeaks (Pepke et al., 2009). Cisgenome identifies peaks separately on the sense and antisense strands and then identifies the binding region between the two peaks.

SiSSRs performs peak-calling in a defined window. Default size is 20 base pairs (bp), however, users can set their own window size. SiSSRs usually calls a higher number of binding sites than other methods (Jothi et al., 2008). When two peaks are very close to each other, depending on the algorithm and parameters used, one peak-caller tool could combine the two peaks and report it as one, while another tool could report it as two peaks. This could affect downstream analysis.

A typical final output of all of these tools is a genomic region tab delimited text file with at least 3 columns, chromosome, start and end coordinates (explained in Section 1.8.7), however, most tools provide additional information either as additional columns or in another file. Table 1.2 summarises different peak-calling tools based on the peak criteria, tag shift functionality, control data handling, false discovery rate (FDR), user input parameters and strand-based artifact filtering.

	Profile	Peak criteria ^a	Tag shift	Control data ^b	Rank by	FDR ^c	User input parameters ^d	Artifact filtering: strand-based/duplicate ^e	Refs
CisGenome v1.1	Strand-specific window scan	1: Number of reads in window 2: Number of ChIP reads minus control reads in window	Average for highest ranking peak pairs	Conditional binomial used to estimate FDR	Number of reads under peak	1: Negative binomial 2: conditional binomial	Target FDR, optional window width, window interval	Yes / Yes	(Ji et al., 2008)
ERANGE v3.1	Tag aggregation	1: Height cutoff High quality peak estimate, per-region estimate, or input	High quality peak estimate, per-region estimate, or input	Used to calculate fold enrichment and optionally <i>P</i> values	<i>P</i> value	1: None 2: # control / # ChIP	Optional peak height, ratio to background	Yes / No	(Johnson et al., 2007; Mortazavi et al., 2008)
FindPeaks v3.1.9.2	Aggregation of overlapped tags	Height threshold	Input or estimated	NA	Number of reads under peak	1: Monte Carlo simulation 2: NA	Minimum peak height, subpeak valley depth	Yes / Yes	(Fejes et al., 2008)
F-Seq v1.82	Kernel density estimation (KDE)	<i>s</i> s.d. above KDE for 1: random background, 2: control	Input or estimated	KDE for local background	Peak height	1: None 2: None	Threshold s.d. value, KDE bandwidth	No / No	(Boyle et al., 2008)
GLITR	Aggregation of overlapped tags	Classification by height and relative enrichment	User input tag extension	Multiply sampled to estimate background class values	Peak height and fold enrichment	2: # control / # ChIP	Target FDR, number nearest neighbors for clustering	No / No	(Tuteja et al., 2009)
MACS v1.3.5	Tags shifted then window scan	Local region Poisson <i>P</i> value	Estimate from high quality peak pairs	Used for Poisson fit when available	<i>P</i> value	1: None 2: # control / # ChIP	<i>P</i> -value threshold, tag length, mfold for shift estimate	No / Yes	(Zhang et al., 2008)
PeakSeq	Extended tag aggregation	Local region binomial <i>P</i> value	Input tag extension length	Used for significance of sample enrichment with binomial distribution	<i>q</i> value	1: Poisson background assumption 2: From binomial for sample	Target FDR	No / No	(Rozowsky et al., 2009)

						plus control			
QuEST v2.3	Kernel density estimation	2: Height threshold, background ratio	Mode of local shifts that maximize strand cross-correlation	KDE for enrichment and empirical FDR estimation	q value	1: NA 2: # control / # ChIP as a function of profile threshold	KDE bandwidth, peak height, subpeak valley depth, ratio to background	Yes / Yes	(Valouev et al., 2008)
SICERv1.02	Window scan with gaps allowed	P value from random background model, enrichment relative to control	Input	Linearly rescaled for candidate peak rejection and P values	q value	1: None 2: From Poisson P values	Window length, gap size, FDR (with control) or E -value (no control)	No / Yes	(Zang et al., 2009)
SiSSRs v1.4	Window scan	$N_+ - N_-$ sign change, $N_+ + N_-$ threshold in region ^f	Average nearest paired tag distance	Used to compute fold-enrichment distribution	P value	1: Poisson 2: control distribution	1: FDR _{1,2} : $N_+ + N_-$ threshold	Yes / Yes	(Jothi et al., 2008)
spp v1.0	Strand specific window scan	Poisson P value (paired peaks only)	Maximal strand cross-correlation	Subtracted before peak calling	P value	1: Monte Carlo simulation 2: # control / # ChIP	Ratio to background	Yes / No	(Kharchenko et al., 2008)
USeq v4.2	Window scan	Binomial P value	Estimated or user specified	Subtracted before peak calling	q value	1, 2: binomial 2: # control / # ChIP	Target F		

Table 1-2: Publicly available ChIP-seq peak-calling software packages. Modified from Pepke et al. 2009. (Pepke et al., 2009)

^a The labels 1: and 2: refer to one-sample and two-sample experiments, respectively.

^b These descriptions are intended to give a rough idea of how control data is used by the software. 'NA' means that control data are not handled.

^c Description of how FDR is or optionally may be computed. 'None' indicates an FDR is not computed, but the experimental data may still be analyzed; 'NA' indicates the experimental setup (1 sample or 2) is not yet handled by the software. # control / # ChIP, number of peaks called with control (or some portion thereof) and sample reversed.

^d The lists of 'user input parameters' for each program are not exhaustive but rather comprise a subset of greatest interest to new users.

^e 'Strand-based' artifact filtering rejects peaks if the strand-specific distributions of reads do not conform to expectation, for example by exhibiting extreme bias of tag populations for one strand or the other in a region. 'Duplicate' filtering refers to removal of duplicate reads at the same genomic location.

^f N_+ and N_- are the numbers of positive and negative strand reads, respectively.

1.8.7. Genomic Regions/Intervals

The typical minimal output of the peak-calling process is genomic data in the form of chromosome, start and end coordinates which are usually referred as genomic intervals or genomic regions. Many genomic features such as, genes, exon, introns, coding sequences (CDS) are also represented by genomic regions in a similar format. For computational processing of such genomic regions, BED (Browser Extensible Data) and GFF (General Feature Format) file formats are mostly used. Both of these formats are based on tab-delimited text files where each line in the file defines a genomic feature. Generally for transcription factors peaks can range from few thousand bases to tens of thousands bases (~ 5kb - 30 kb) however for histone modification usually a greater number of peaks are called which could be more than 100 kb. More about these formats is discussed in Chapter 2.

1.8.8. Analysis of Enriched Genomic Regions

The peak-calling process, described above, identifies the genomic regions where a transcription factor binds or epigenetic marks exist. Once genomic regions that represent the binding site for transcription factor or histone marks are identified, then we can perform a number of other analysis such as identifying transcription co-factors or pathway/genes that are regulated. Sequence motif analyses can reveal a degree of affinity on various DNA sequences (Kasowski et al., 2010; Hu et al., 2010; Arbiza et al., 2013). These analyses are explained below.

1.8.8.1. Motif analysis

Motifs are short, recurring patterns in DNA sequence that are acknowledged to have a biological function. As they indicate sequence-specific binding sites for transcription factors, therefore, motif-based analysis is used to identify sequence motifs in the binding regions. To initiate the motif analysis, genomic coordinates converted into genomic

sequences usually in FASTA format are required. FASTA is a text based file format in which each nucleotide is represented by its letter (e.g. A for Adenine) (Pearson, 1994). The sequences can be extracted from UCSC Genome Browser or Galaxy and then motif discovery can be performed by any of the motif discovery tools (Table 1.3). Some analysis tools such as HOMER come with their own genomic databases and a genomic region file can be input without the need to manually convert it into FASTA. Some motif analysis tools such as MEME-ChIP and peak-motifs, are part of analysis pipelines that perform several motif analysis steps, while several (e.g. FIMO, HOMER and PATSER) can perform motif prediction and mapping to identify candidate binding sites.

Motif analysis gives insight into the regulatory mechanisms of the factor under study and if a predicted binding motif for that transcription factor is already known, finding the centrally located motif in a genomic region dataset validates the success of the ChIP-Seq experiment (Bailey et al., 2013). Motif analysis also has the power to identify motifs of other proteins, if they are present in the binding regions, that are associated with the transcription factor under study. *De novo* motif discovery analysis identifies novel regulatory sequences and help define their roles.

category	Software tool	Web server	Obtain peak regions	Motif discovery	Motif comparison	Central motif enrichment analysis	Local motif enrichment analysis	Motif spacing analysis	Motif prediction / mapping	Ref.
Motif discovery + more	ChIPMunk	X		X						(Kulakovskiy et al., 2010)
	CisGenome			X	X					(Ji et al., 2011)
	CompleteMOTIFS	X		X	X					(Kuttippurathu et al., 2011)
	MEME-ChIP	X		X	X	X				(Machanick and Bailey, 2011)
	peak-motifs	X		X	X				X	(Thomas-Chollier et al., 2012)
	HOMER		X	X					X	(Heinz et al., 2010)
	Cistrome	X	X	X			X	X	X	(Liu et al., 2012)
Motif comparison	STAMP	X			X					(Mahony and Benos, 2007)
	TOMTOM	X			X					(Gupta et al., 2007)
Motif enrichment/spacing	CentriMo	X				X	X			(Bailey and Machanick, 2012)
	SpaMo	X						X		(Whittington et al., 2011)
Motif prediction/mapping	FIMO	X							X	(Grant et al., 2011)
	PATSER	X							X	(Hertz and Stormo, 1999)

Table 1-3: Software tools for motif analysis of ChIP-seq peaks and their uses. Modified from Bailey et al. 2013 (Bailey et al., 2013).

1.8.8.2. Analysis of Overlapping Binding Regions Reveals Co-operation between Binding Factors

Identifying overlaps in genomic features such as histone modifications and other transcription factor binding sites is a fundamental task in this research (Meyer et al., 2012). Recent studies have revealed that there are often overlaps and co-association between transcription factors at binding sites (Gerstein et al., 2012). Identifying genomic localization of common binding regions is an important biological research question. For example, Motallebipour et al. (Motallebipour et al., 2009a) mapped the DNA binding sites of three important forkhead transcription factors FOXA1, FOXA2, and FOXA3 in human liver hepatocellular cells (HepG2). By comparing the data the study established that FOXA2 interacts with FOXA1 and FOXA3, however, FOXA1 and FOXA3 do not interact. In some studies it is important to identify what binding sites do not overlap to understand the regulatory mechanism for example Schmidt et al. (Schmidt et al., 2010) performed ChIP-Seq experiments on Cohesin proteins (RD21, STAG1, SA1), CTCF and ER α in human breast cancer cells (MCF-7) to identify that Cohesin regulates gene expression in a tissue-specific manner, independent of CTCF binding.

Once enriched genomic region datasets are obtained from the peak-calling process, these types of overlap analyses can be performed by software tools that can perform set-like operations using the genomic coordinates of the binding regions, for example BEDTools (Quinlan and Hall, 2010a), Pybedtools (Dale et al., 2011b), GenomicTools (Tsirigos et al., 2012), BEDOPS Tools (Neph et al., 2012). These are command line tools that are primarily designed to run on Linux/Mac environment. They can compare two genomic regions files and can report overlapping or non-overlapping regions. A user needs to learn about the use of their various parameters that can be employed in different situations. BEDTools, Pybedtools and GenomicTools load all of the data in computer memory and perform sorting and indexing

in memory, therefore fail on larger files. On the contrary, BEDOPS provides a separate utility (tool) to sort the files first and then loads only the information required to calculate next line of output, keeping memory utilisation and run time to smallest level (Neph et al., 2012).

Pybedtools is a Python (computer language) interface for BEDTools so essentially overlap analysis options are similar to what BEDTools offers, however, the Python interface makes integration with the Python language and writing sophisticated queries easy. Similarly GROK (Genomic Region Operation Kit) and GenomicTools provide C++ API (application programming interface) to C/C++ programmers and claim a better efficiency in terms of time and memory requirements. GROK can also be integrated with R programming language. All of these tools are command line, whereas, Cisgenome and Galaxy provide a graphical interface, however, the analysis options are very basic. The UCSC table browser (Karolchik et al., 2004) provides browser-based function of genomic regions, however, the input is restricted to 1,000 regions (Zammataro et al., 2014).

All of the tools in Table 1.4 lack a charting tool such as drawing a Venn diagram to graphically represent the level of overlap of two datasets.

Tool	Features	GUI	Query interface *	Statistical Significance	Reference
BedTools	++++				(Quinlan and Hall, 2010a)
PyBed Tools	+++++		✓		(Dale et al., 2011b)
GROK	+++++		✓		(Ovaska et al., 2013)
BEDOPS	++++				(Neph et al., 2012)
Cisgenome	+	✓			(Ji et al., 2008)
GenomicTools	++++		✓		(Tsirigos et al., 2012)
Galaxy	+	✓			(Cock et al., 2013)
MULTOVL	++++			✓	(Aszodi, 2012)
UCSC Table Browser	++	✓			(Kent et al., 2002a)

Table 1-4: Comparison of tools that operate on genomic regions.

GUI = Graphical User Interface

+ = Very basic overlapping features, +++++ = Sophisticated overlapping features

* Easy integration with programming languages

Various peak caller software call different number of peaks (genomic regions) depending on the algorithm and parameters in use for a peak-caller tool, therefore, it is required that when studying the degree of overlap of two datasets it should be identified whether the two datasets overlap by chance or the overlap is statistically significant. Table 1.5 lists recently published algorithms for determining the significance of overlap of two datasets. Out of all of the tools listed in Table 1.4 and Table 1.5 only MULTOVL operate on genomic regions and also computes the statistical significance of overlapping regions. Whereas only the IntervalStats (Chikina and Troyanskaya, 2012) tool provides a p-value for each region. More details about the IntervalStats implementation is discussed in Chapter 2.

Tool	Genomic region operation	Simulation	p-value for each region	Ref.
MULTOVL	✓	✓		(Aszodi, 2012)
IntervalStats			✓	(Chikina and Troyanskaya, 2012)
Binary Interval Search (BiTS)		✓		(Layer et al., 2013)
Genometricorr		✓		(Favorov et al., 2012)

Table 1-5: List of tools that compute the statistical significance of overlapping regions.

1.8.8.3. Annotation analysis

Annotation of genomic regions is another fundamental task in studying gene regulation by transcription factors (Meyer et al., 2012). This in essence refers to the association of genomic region data with information about nearby genomic features. Gene annotations including gene identities, gene names, chromosome, strand, coordinates of transcription start site (TSS), end site (TES), coding sequence (CDS) and exon positions can be downloaded from various public servers such as UCSC Genome Browser (Kent et al., 2002a) or Ensembl (Hubbard et al., 2002) as BED, GFF or other tab-delimited text format. Genomic regions are then compared to annotation to perform various analyses such as identifying the closest genes regulated by the genomic regions, regions that are within certain base pairs away from TSS or TES, distance of genomic regions from nearest TSSs. Tools listed in Table 1.4 can also be used to perform these analyses, however, out of these tools Cisgenome and UCSC Table Browser provide a user-friendly graphical interface.

1.9. Challenges / Gaps

Analysis of enriched genomic regions is a complex process due to the fact that current genome-wide approaches produce very large datasets, often containing tens of thousands of genomic regions, studies often seek to compare several experimental conditions, and results are highly dependent on the software tools employed and the analysis parameters used. In addition, most analyses are carried out through several steps, requiring the use of multiple software tools for specific tasks such as genomic alignment, binding region identification, motif analysis and gene annotation, as discussed above. Publicly deposited datasets are rapidly expanding in size and complexity. However there is currently a lack of tools that curate genomic regions, therefore, it is now acknowledged that genomics studies need more user-friendly and sophisticated data analysis and interpretation tools (Barski and Zhao, 2009; Dale et al., 2011b; Krystkowiak et al., 2013). One of the most challenging steps in genomic analysis is to compare multiple genomic region datasets from various experimental procedures, tissue type, temporal or developmental stages (Taslim et al., 2009; Sandmann et al., 2006). When comparing peaks, investigating simply the overlap of two sets of peaks may not represent the optimal approach (Bailey et al., 2013) unless various experimental conditions, tissue types and statistical significance are taken into account. Bench biologists often experience the following challenges:

1. Datasets from previous publications reporting transcription factor binding sites or histone modifications are scattered on journal websites, Gene Expression Omnibus (GEO) and other public servers and there is no central public resource or tool available where investigators can easily select datasets based on tissue type and conditions of an experiment.
2. There are a few tools that provide a user-friendly interface to study overlapping or non-overlapping datasets, however, they are very limited in features and

sophistication. For example, the UCSC table browser provides the option to extract the data that overlap with features of another selected table, however, the tool does not report the total number of base pairs overlapping with other regions or distance from centre of the two comparing sets. Similarly Galaxy and Cisgenome provide very limited options to study overlapping genomic regions. However, they provide limited functionality for the user to set overlap criteria.

3. Most tools do not provide for each overlap found all the regions involved in that overlap, the overlapping sections, total overlapping base pairs, distance from centres of regions. Most tools lack options to determine statistical significance of overlap results.
4. There are no genomic region analysis tools that combine a curated database of published genomic region data with tools to identify and analyse overlaps between genomic datasets.

The existence of a wealth of published data sets now presents unprecedented opportunities for data mining in large databases of archived genomic region data.

1.10. Aims

Medical research depends heavily on the advent of statistical, mathematical and computer science algorithms and software. However biologists often lack expertise in computing and mathematics, which limits their ability to exploit high throughput genomic approaches. In contrast, computational scientists need a grounding in biology to write tools that can address meaningful questions in genomics. Therefore to address these needs this thesis had the following aims:

- i. This project aims to bridge this gap by providing easy Graphical User Interface (GUI) tools with illustrated use of the latest genome analysis tools.

- ii. Collection of publicly available genomic region datasets from published ChIP-Seq and ChIP-chip studies into a single searchable curated repository.
- iii. Annotation of the genomic regions by organism, reference assembly, cell line, factor, conditions, total reported regions, peak caller software, experiment type (ChIP-chip/seq), author and publication year.
- iv. Development of user-friendly genomic software for Windows as a standalone desktop application. The desktop version for bench biologists will be able to archive unlimited data in a personalised database and will provide easy tools to analyse the transcription factor binding sites and histone modifications datasets.
- v. Development of database resource and tools for large bioinformatics facilities that operate servers on Unix, Linux or Macintosh operating systems.
- vi. Development of a project website with installation guidelines and options to download the software for various computer platforms.
- vii. Analysis of the collected data that will serve as an example for further genomic research using the resource.

Chapter 2: Selection of suitable computational resources to build a genomic database

2.1. Introduction

Recent publications have highlighted the complexity of transcriptional regulation, and it is now known that there are many factors that must be analysed in parallel to interpret complex biological processes as explained in the Transcription Factors and the Epigenetics sections in Chapter 1. One of the data analysis outcomes of such genomic studies are DNA target regions which are generally referred as genomic regions or genomic intervals. Genomic regions are datasets containing locus information per chromosome (as explained in Section 1.5.6 and 2.7). These data need to be stored in a way that enable application of mathematical operations to them. The data should be stored and retrieved by a technology that can be easily accessible to biologists, preferably at no cost and there should be a large community that make use of the technology. Tools or resources written in a technology well used in a community of users will be more likely to be used and maintained by the community. However, at the same time a developer of the resource/tool should not exclude other technologies which could provide future benefits and prospects. Therefore I began with an extensive survey of available platforms and technologies that were used in genomic research. In order to do this I collected a number of datasets first to understand what sort of the data the technology needs to handle. Later in the chapter I explained various issues, database benchmarking and testing of the software that I developed during my candidature. We named this resource and its tools the Binding Site Analyser (BiSA) to analyse, annotate and interpret genomic regions statistically and graphically.

2.2. Collection of Datasets

PubMed and Google Scholar were searched for ChIP-chip and ChIP-seq transcriptional regulation studies describing transcription factor binding sites and histone modifications and

it was identified that much of the datasets were deposited in public repositories such as Gene Expression Omnibus (GEO) (Edgar et al., 2002), European Bioinformatics Institute (EBI, <http://www.ebi.ac.uk/>), Cistrome Project (Liu et al., 2011a). Some datasets were collected from journal websites or directly from the authors. The source of the data and additional comments, if there were any, were also recorded.

~1000 datasets of transcription factor binding sites and histone modifications were downloaded. The datasets were annotated with information about factor name, cell line, peak-caller software, experiment type and sample treatment by reading the papers and associated files. The total collection comprises ~24 million genomic regions with a total combined size of about 1 Gigabyte.

2.3. Choice of Operating Systems

It is very important to decide the operating system on which a tool would run and this affects the downstream decision of the database and the language of development. I reviewed two major types of operating systems, i) Unix-like operating systems, ii) and Microsoft Windows. Unix-like operating systems include variants of Linux and Apple Macintosh (Mac) OS X and later (Katayama et al., 2000; Winterbottom and Wilkinson, 1990; Accetta et al., 1986; Meyers and Lee, 2011). Unix-based operating systems are very popular in the field of bioinformatics research and numerous tools are written to exclusively run on these operating systems (Dudley and Butte, 2009; Stajich and Lapp, 2006a). The popularity of these operating systems is credited for the range of free open source languages available and many open source projects (Oinn et al., 2004; Novak et al., 2013; Blankenberg et al., 2010; Giardine et al., 2005).

On the other hand Microsoft Windows is the dominant operating system and holds ~90% of the operating system market (Share, 2013) and is widely known to be user-friendly. Many

researchers would have experience working on Windows. Therefore, initially I built BiSA for Windows operating system as a desktop application. However, identifying limitation with scalability as explained in the section 3.3, a web-based version for the Unix-like operating system was developed.

For BiSA for Unix-like operating systems, we chose Red Hat Linux because it is one of the most popular distributions of Linux and there is a large community of users (McCarty, 2004) and is the platform recommended by the Sydney University Information and Communication Team where I conducted my research.

2.4. Database Selection

The recent revolution in whole genome census approaches has seen an exponential increase in available data sets describing genomic features, such as transcription factor binding sites and histone modifications. Manual management of such files for curation and identifying relationships is cumbersome (Meyer et al., 2012) . Database systems are the tools that have been utilised in many fields to ease the task of handling data and their relationships. One of the main aims of this project was to develop a user-friendly archival and retrieval system for genomic regions coupled with tools to analyse data using database systems. Therefore various database systems are reviewed to identify the best database that could handle genomic regions data efficiently as explained below.

2.5. Relational or Non-Relational Databases

There are two broad choices in database systems, relational and non-relational databases. For the last 4 decades classical databases have been built on the relational database management systems (RDBMS) model. A RDBMS stores data in a strict structure known as a schema. The schema comprises tables, views and stored procedures and each table has a common structure for all of its rows. These databases support a standard query language known as SQL

(Structured Query Language), therefore, today such databases are commonly referred to as SQL databases. In recent years non-relational databases, referred to as NoSQL, have been developed to store non-structured or schema-less data by reducing strict data (Kala Karun and Subu, 2013). NoSQL databases claim high performance, availability and easy scalability and replication over many servers in comparison to SQL databases. In NoSQL databases, such as MongoDB, instead of tables, databases have collections. NoSQL databases employ various non-relational techniques to store data: for example, Apache Cassandra stores data as key/value stores, MongoDB stores data as a collection of documents, AllegroGraph or Neo4J are graph databases (Have and Jensen, 2013; Marin and Dragos, 2013). Therefore NoSQL databases provide flexibility in schema design in order to deal with the challenge of scalability with SQL databases. Table 2.1 provides an overview of strengths and weaknesses of the two types of database models.

Apart from a powerful query language, SQL databases have four other strengths i) all data operations (transactions) will commit (permanent) or not at all, ii) ensuring the complete change of state of the database, iii) ensuring independent running of transactions and iv) ensuring the state of completed transactions will persist (Cattell, 2011; Leavitt, 2010; Fortier et al., 1994). However, critics argue that the above constraints affect the performance of SQL databases; therefore, with minimum constraints NoSQL databases offer better speed, efficiency and scalability. With a powerful query language/interface SQL databases are a good choice for databases which are not expected to grow out of the capacity of a server, since scaling and distribution of SQL databases are not easy.

	Strengths	Weaknesses
Relational databases (SQL)	<ul style="list-style-type: none"> • In common use • Built-in powerful query language • Best suited to structured data • Retrieval based on complex query is easy • Data operations precise 	<ul style="list-style-type: none"> • Not easy to scale • Less efficient on big data • Scaling is not easy • Distribution of data across partitions or servers is not easy
Non- Relational databases (NoSQL)	<ul style="list-style-type: none"> • Easy to scale • Efficient handling of big data • Lower administration • Schema-free so needs less development time • Better handling of unstructured data • Flexible data models 	<ul style="list-style-type: none"> • Lack of support • Lack of reporting tools • Lack of standardisation • Complex coding to retrieve data

Table 2-1: Comparison of relational (SQL) databases and non- relational (NoSQL) databases.

Keeping in mind the features of the two database types, it was clear that we were not dealing with the dimensions of data that NoSQL databases are designed for. Just to illustrate the data sizes the NoSQL databases are designed to deal with: Google web-indexing, the world's largest search engine, uses Bigtable (Chang et al., 2008), Amazon is handling its billions of transactions by Dynamo and Cassandra is being used at Facebook inbox search (Kala Karun and Subu, 2013). I have also observed that NoSQL databases are not often used in bioinformatics (Have and Jensen, 2013). For example Galaxy, web-based bioinformatics

analysis platform, uses SQLite by default, UCSC genome browser and Ensemble use a MySQL database (Hermida et al., 2013; Kent et al., 2002a). Recently Paila et. al. have written a database application GIMINI for genetic variation and genome annotations and they preferred a relational database (SQLite) because of the native support for powerful SQL data exploration queries and its familiarity to many researchers (Paila et al., 2013).

We decided on the selection of SQL versus NoSQL database-type based on the following requirements for our application needs:

1. In first round of development BiSA for Windows will be a desktop personal database resource and typically will be serving only one user at a time.
2. In second round of development, BiSA web-based version unlike some of the commercial web servers, is going to be used by a limited number of investigators.
3. Transcription factor binding sites and histone modifications are the final outcome of a ChIP-Seq experiment so data is usually refined and small at this stage. The file size for transcription factor binding sites is usually just a few hundred kilobytes and histone modifications or genetic annotations are less than a few hundred megabytes. The total size of our collected ~1000 datasets was slightly less than 1 Gigabyte (GB).
4. The data is usually available as tab delimited text files and the format is standardised (explained in the Section 2.6).
5. Most researchers and computational biologists are familiar with relational databases, and this is likely to increase the adoption of BiSA.

Therefore considering all of our requirements I decided to choose a relational (SQL) database-type for the development of BiSA.

Having decided on SQL, there are a number of SQL databases available, so I needed to work out which one would best be suited to our requirements. Therefore, I also performed an

extensive review of the databases with the criteria as being supported on Windows and Unix-like operating system.


2.6. Reviewing Relational (SQL) Databases







I reviewed the seven most widely used SQL databases (db-engines.com) for the suitability of BiSA. Table 2.2 summarises the four important features (operating system support, who owns or maintains the database, the type of license offered and maximum database size supported by a database engine) of seven well-known and widely used databases.

SQLite is a serverless file-based database engine and the default database for many bioinformatics tools (Paila et al., 2013), however, SQLite is maintained by only three developers and is a relatively young database (Hipp et al., 2013). Another shortcoming is that unlike a dedicated database server SQLite is not designed to deal with the challenge of handling multiple jobs at a time and transactional locks may occur on the production server. These locks cause timeouts and job errors. On this basis, I eliminated SQLite for consideration.

Four databases, Microsoft SQL Server (abbreviated as SQL Server), Microsoft Access, FileMaker and Oracle are proprietary, and consequently there is a cost associated with their utilisation (Table 2.2). Microsoft Access and FileMaker are used for small business applications such as invoicing, inventory control, or basic business applications. The maximum database size of Microsoft Access is 2 Gigabytes, and aimed at very small business applications, therefore I did not consider it any further as it would be limited in terms of extendability. FileMaker is a more professional database software than Microsoft Access with its cross compatibility on Windows and Mac operating systems and it offers an Advanced Server version as well. However, firstly it is a costly application, and secondly it is designed to develop business solutions and is not considered a general purpose database.

Microsoft SQL Server and Oracle are also proprietary databases and not free of cost, however, both products provide free Express Editions. In Express Editions, the memory utilisation is limited to 1GB though and only 1 central processing unit (CPU) is supported. SQL Server Express Edition supports maximum 10 GB while Oracle Express Edition supports 11 GB databases. These size limitations seem reasonable for the type of datasets we have been dealing with. In contrast to Oracle, many educational and research organisations have enterprise agreements for Microsoft products such as Microsoft Office and SQL Servers professional editions. Therefore free access to a professional or standard Microsoft SQL Server for a researcher is more likely to be available. Students can also download a free copy of SQL Servers from DreamSpark.com. SQL Servers professional editions can handle database sizes upto 534 Peta bytes. Moreover Microsoft SQL Server is natively supported on Windows and other Microsoft languages. Therefore Microsoft SQL Server seems a reasonable choice for BiSA on the Windows platform. However, there is a disadvantage of this choice in that there is no Unix-like version available. Therefore, for BiSA for Unix-like operating systems MySQL or PostgreSQL seem good candidates to be considered. Both MySQL and PostgreSQL have Windows versions available so they could be potentially used in both operating environment. These databases are widely used in a range of business applications and in research, therefore, there is large community-base and extensive documentation is also available (Welling and Thomson, 2003; He-qin, 2007; Di Giacomo, 2005; Paulson, 2004).

Databases	Operating System	Maintained By	License	Maximum Database Size
SQL Server		Microsoft	Proprietary	524 Petabytes ¹

Microsoft Access		Microsoft	Proprietary	2 Gigabytes
Oracle		Oracle Corporation	Proprietary	Unlimited ²
MySQL		Oracle Corporation	GPL (General Public License) or Proprietary	Unlimited
PostgreSQL		Global Community	PostgreSQL open source license	Unlimited
SQLite		The SQLite Development Team	Public domain	128 Terabyte
FileMaker		FileMaker Inc.	Proprietary	8 Terabyte

=Windows, =Linux, =Apple Macintosh (Mac)

Table 2-2: Comparison of databases for operating system support, maintained by, license-type and maximum database size of seven top databases.

¹ Free Microsoft SQL Express version's maximum storage capacity is 10GB.

² Free Oracle Express version's maximum storage capacity is 11GB.

2.7. Development of Overlapping Regions Algorithms

From the initial analysis I short-listed three database Microsoft SQL Server, MySQL and PostgreSQL. To identify which database would be the best for development of a genomic region database, I developed and published (Khushi, 2015) a novel the RegMap (Region Mapping) benchmarking algorithm that operates on genomic locations natively in the database. Since previously developed algorithms act on flat files and at the time there was no

published algorithm available that analyse genomic regions natively in a database system. Using RegMap I tested the performance of these candidate databases. Following sections explain the various aspects of RegMap algorithm.

2.7.1. Processing Genomic Regions

Many genomic features are saved in data repositories as tab-delimited text files. For computational processing of such files, BED (Browser Extensible Data) and GFF (General Feature Format) file formats have been widely used. BED format defines the first three fields as required and nine additional optional fields. The first three required fields are tab-delimited chromosome, start and end while strand as plus (+) or minus (-) sign is saved in the sixth optional field. GFF format defines nine required fields where chromosome or name of the feature is saved in the third column and fourth and fifth field contain start and end of the genomic position. Each line in the file defines a complete genomic feature. The other important difference between the two file formats is the start index of the first base in a chromosome. The BED format defines the chromosome start as zero, on the other hand, in GFF the first base is numbered 1. Therefore, region length in GFF file is calculated by subtracting the start from the end coordinate and increasing the result by 1, for example, for genomic region starting at base 100 and ending at base 200, the region length is calculated as:

$$\textit{Region length} = (\textit{end} - \textit{start}) + 1 = (200 - 100) + 1 = 101$$

However the same interval in BED format will be expressed as starting from 99 and ending at 200 so the region length if the feature is saved as BED format will be calculated as:

$$\textit{Region length} = (\textit{end} - \textit{start}) = 200 - 99 = 101$$

This shows that the BED format requires one less mathematical operation and so is more computationally efficient. Therefore we decided to save all start coordinates zero indexed as

in the BED file format to keep mathematical operations efficient. When a dataset from the GFF file format is imported its start coordinate was decreased by 1 before saving into the BiSA database.

Two genomic intervals or regions are said to often intersect or overlap if both intervals share at least one base pair in common on a same chromosome. However, this can be user-defined, for example, intervals may be said to overlap only if they share 100 bp or only if their centres are within 50 bp.

To illustrate further, consider two regions Chr1:10-20 and Chr1:18-30, these regions are said to overlap by 3 bases on 18,19 and 20 base pair positions on chromosome 1. So we devised a variable 'bp overlap' which is calculated to be positive by counting the number of base pairs in common between two regions or negative when calculating distance between the ends of two non-overlapping regions.

2.7.2. Calculation of Base Pair Overlap (bp overlap)

The bp overlap is represented as a positive number for regions having bases in common between two sets (shown as shaded region in Figure 2.1), and for non-overlapping regions the 'bp overlap' is shown as a negative number of bases away from the corresponding ends. To illustrate the algorithm let us consider two regions A and B in the three scenarios (Figure 2.1).

- i) One region is completely within the other or complete overlap. In this case the bp overlap is simply a positive number reported by calculating the region length of the smaller region that lies within the other region. If the two regions completely overlap each other then the length of either region can be reported as bp overlap. Since the start coordinate is saved in the BED format therefore region length can be calculated by just subtracting the start from the end coordinate as explained in the section above.

Computationally this case (Figure 2.1 i-a) is identified by checking if 'A.End' is less than or equals 'B.End' and 'A.Start' is greater than or equals 'B.Start', implies A is within B, calculate region length of A. SQL pseudocode extract given below:

```
WHEN A.End <= B.END AND A.Start >= B.Start
      THEN (A.End - A.Start)
```

To identify the second case (Figure 2.1 i-b), if region B lies within region A, we need to verify if 'B.End' is less than or equals 'A.End' and 'B.Start' is greater than or equals 'A.Sart', calculate the region length of B.

```
WHEN B.End <= A.END AND B.Start >= A.Start
      THEN (B.End - B.Start)
```

- ii) The second scenario is when region A lies on the left side of region B (Figure 2.1-ii). In this situation the two regions could share bases in common or could be distant to each other. Computationally this is checked if 'A.End' is less than or equals 'B.End' and 'A.Start' is less than or equals 'B.Start'. The bp overlap is calculated by subtracting 'B.Start' from 'A.End'.

```
WHEN A.End <= B.End AND A.Start <= B.Start
      THEN (A.End - B.Start)
```

In the above calculation, for the first situation (Figure 2.1 ii-a) bp overlap will be reported as a positive integer. Whereas, for the second situation (Figure 2.1 ii-b) when there are no common bases in the two regions, the bp overlap is calculated as a negative number. This is because in this case the 'B.Start' coordinate is greater than 'A.End' therefore $(A.End - B.Start)$ will be a negative number.

- iii) In the third case region A could be on the right side of region B. As above, the two regions could overlap or can be apart without intersecting each other. This is ensured

by checking that if 'A.End' is greater than or equals 'B.End' and 'A.Start' is greater than or equals 'B.Start'. The bp overlap is calculated by subtracting 'A.Start' from 'B.End'.

```
WHEN A.End >= B.End AND A.Start >= B.Start
    THEN (B.End - A.Start)
```

Similar to the above case the overlapping regions (Figure 2.1 ii-a) will have positive bp overlap and non-overlapping regions will have negative bp overlap (Figure 2.1 ii-b).

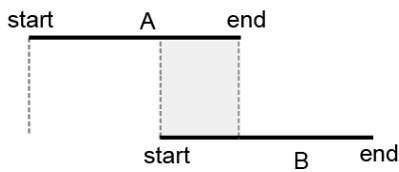
i-a)



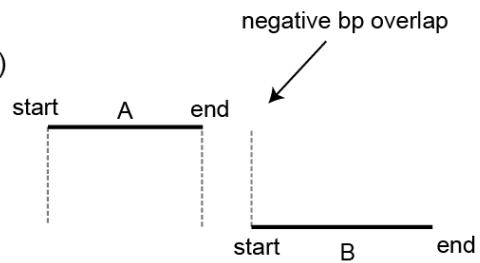
i-b)



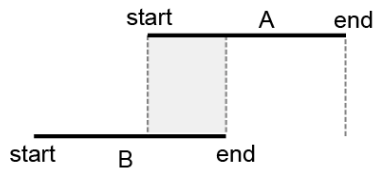
ii-a)



ii-b)



iii-a)



iii-b)

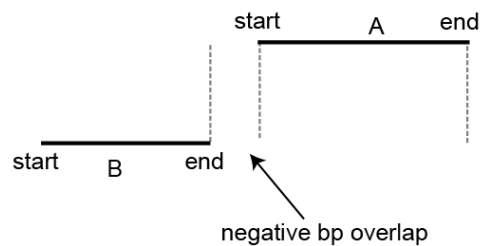


Figure 2-1: Various relative positions of the two regions. i) One region is completely within the other. ii) The overlapping or non-overlapping region A is on the left side of region B. iii) The overlapping or non-overlapping region A is on the right side of region B.

2.7.3. Calculation of Distance between Region Centres

We also calculate the distance between the centres of the two nearest regions, we refer to this as ‘centre distance’. Some overlap analyses of transcription factors require restricting the overlap analysis based on maximum centre distance. This is important because transcription factor binding sites are typically 100-200 bases, however, some regions reported as binding sites during the peak-calling process are many thousand bases long due to multiple consecutive peaks merged and reported as one peak as explained in the peak-calling Section 1.5.5. Therefore a researcher might like to restrict analysis based on distance from region centres.

To calculate the centre distance, the centres of the two regions are calculated and then the distance between the centres is calculated by the following SQL code.

$$(A.End + A.Start)/2 - (B.End + B.Start)/2$$

Since the distance can be negative if the location of the centre of the region B is greater than location of the centre of region A, therefore, an ABS SQL functions is used to return the absolute (positive) value of the calculation by ignoring the negative sign.

$$ABS (A.End + A.Start)/2 - (B.End + B.Start)/2$$

The complete SQL code that implements the above explained ‘bp overlap’ and ‘centre distance’ algorithm is given in the Box 2.1.

The code (Box 2.1) also contains information about the two regions and is saved as a *view* in the database, named *vwKBCompareSites*. A view acts like a table and performs all the

calculations during the run-time. Therefore we can run a query against the *view* and can also restrict results based on bp overlap, centre distance or chromosome.

```
select A.Chr Chr_A, A.start Start_A, A.[end] End_A,
       B.Chr Chr_B, B.start Start_B, B.[end] End_B,
       ABS((A.[end] + A.start)/2 - (B.[end] + B.start)/2) CentreDistance,
       bpOverlap = CASE
           WHEN A.[end] <= B.[END] AND A.Start >= B.Start
               THEN (A.[End] - A.Start)
           WHEN B.[end] <= A.[END] AND B.Start >= A.Start
               THEN (B.[End] - B.Start)
           WHEN A.[end] <= B.[END] AND A.Start <= B.Start
               THEN (A.[End] - B.Start)
           WHEN A.[end] >= B.[END] AND A.Start >= B.Start
               THEN (B.[End] - A.Start)
           END
from   dbo.KBSites A inner join dbo.KBSites B on A.Chr=B.Chr
```

Box 2.1: *SQL code to calculate the bp overlap and the centre distance. Key words are coloured.*

2.7.4. Extracting Overlapping Sections of Regions Common in Two Datasets

One region of a dataset can overlap two regions in another dataset. Therefore, certain analyses require the identification of overlapping sections of the regions common in two datasets. The overlapping regions must be intersected by at least 1 base pair, shown by grey shaded rectangles in Figure 2.1. Since both compared regions are restricted to be on the same chromosome, chromosome information can be taken from either region. Similarly if start or

end coordinates are the same then they can be taken from either region. Computationally to identify the start of the overlapping section, the two regions are compared and which region's start is greater is checked; the greater start is taken as the start of the overlapping section of the region i.e.

```
WHEN Start_A>Start_B Then Start_A
```

```
WHEN Start_A<Start_B Then Start_B
```

On the other hand, to identify the end of the overlapping section the smaller end coordinate is chosen i.e.

```
WHEN End_A>End_B Then End_B
```

```
WHEN End_A<End_B Then End_A
```

Box 2.2 is an extract of the SQL code that is used to extract the overlapping regions.

```
select distinct Chr_A Chr_common,  
               CASE  
                   WHEN Start_A=Start_B Then Start_A  
                   WHEN Start_A>Start_B Then Start_A  
                   WHEN Start_A<Start_B Then Start_B  
               End as Start_common,  
               CASE  
                   WHEN End_A=End_B Then End_A  
                   WHEN End_A>End_B Then End_B  
                   WHEN End_A<End_B Then End_A  
               End as End_common  
from vwKBCompareSites where bpOverlap>=1
```

Box 2.2: *SQL Code to extract the overlapping sections of the regions common in two datasets.*

2.8. Benchmarking Database Performance using RegMap

Using the above explained algorithm I developed a benchmarking script called RegMap (Region Mapping) and benchmarked the performance of SQL Server, MySQL and PostgreSQL for i) insertion of data ii) identification of overlapping regions. I have also compared the RegMap performance against database built-in spatial functions which, provided very limited functionality.

The time rounded to the nearest second for each operation was recorded. The time noted for insertion of data also included the time it took to generate random regions for each product.

RegMap benchmarking script generated all the required objects in a working database. The algorithm was completely developed in native SQL (Structured Query Language) and is therefore compatible with all SQL databases.

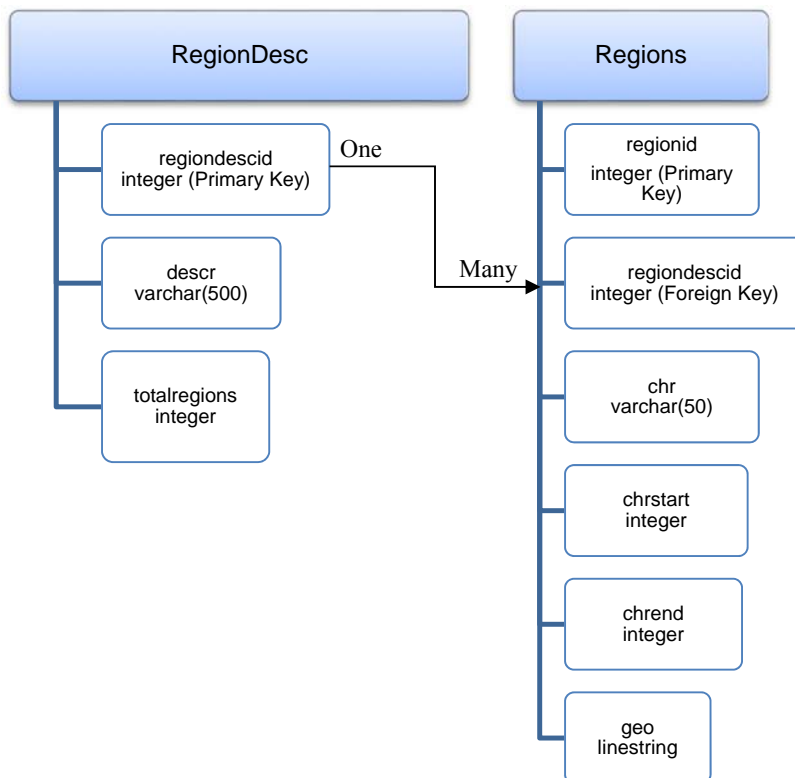


Figure 2-2: Database structure employed to benchmark the performance of database systems.

Chromosomes were saved as character data-type and start and end coordinates as integer data-type (Figure 2-2). Primary key fields automatically generates clustered indexes, apart from this index no other index-type was created. To compare performance with database built-in spatial functions the coordinates were saved as linear spatial data-type, *lseg* in case of PostgreSQL or *linestring* in MySQL. Genomic regions were saved in the *Regions* table and were linked with the *RegionDesc* table where annotation of the regions was saved, thus simulating a production usage. Each region in the *Regions* table was automatically assigned a unique database id (Primary Key). The start coordinates of the genomic regions were indexed from 0, as described above, to speed up calculations, therefore region length was calculated by subtracting the start from the end coordinate.

A total of 1005 datasets of transcription factor binding sites and histone marks from previous publications on human and mouse assemblies were collected. This ‘Knowledge Base’ was used to perform insertion and searching benchmarking.

All testing and benchmarking were performed on PostgreSQL 9.0 and MySQL Community Server 5.6.15 GPL (x86_64) installed on HP Compaq 8200 Elite running Windows 7 Professional having 4 core 2.5 GHz processor with 8GB memory. RegMap code was run in MySQL Workbench 6.1 for MySQL server and in pgAdmin III 1.81 for PostgreSQL benchmarking maintaining the default settings of each database. Client and database servers were on the same computer. MySQL Server supports a number of storage engines, in this performance benchmarking the two most widely used: InnoDB and MyISAM storage engines (Sheldon and Moes, 2005; Padilla and Hawkins, 2011). The results of 100 simulations were averaged for all operations. The default random region size was set to 500, however, this setting can be changed in the script.

I also performed RegMap benchmarking on Microsoft SQL Server, however, the detailed results are suppressed here, because, Microsoft licencing agreement (Microsoft.com) prohibit publication of any benchmarking results.

2.8.1. Benchmarking for Insertion of Data

RegMap randomly generates regions in memory between the specified lower and upper range and then finally saves the data in the database in a single transaction. I identified that this technique was faster for both databases since each time an insert statement is executed against the database there are transaction overheads. So generating and saving regions one by one was much slower.

I tested the performance by generating 5K, 10K, 20K, 40K and 80K regions and identified that PostgreSQL's generation of random regions and insertion was much faster than MySQL in both InnoDB and MyISAM storage engines. MySQL's insertion of regions was dramatically slower and the time taken was almost double by doubling the number of regions inserted (Figure 2.2). MySQL-InnoDB performed slightly better than MySQL-MyISAM, therefore, in Figure 2.2 performance of MySQL-InnoDB is reported. PostgreSQL (RegMap) generated and saved 5K random regions in 1 second as compared to 219 second in MySQL-InnoDB and 237 second in MySQL-MyISAM, indicating that MySQL was ~220 times slower. This difference dramatically increases for a much larger number of regions. For generating 80K random genomic regions PostgreSQL took 4 seconds as compared to 3,596 sec in MySQL-InnoDB and 3,680 second in MySQL-MyISAM.

In addition, the database write performance was tested by importing the 1005 files consisting of 23,827,431 real genomic regions, collected from previous studies, into both databases using bulk import statements of the databases. PostgreSQL *COPY* statement while MySQL *LOAD DATA INFILE* statement was used for this purpose. I performed the import of each file in three steps: i) data was imported into a staging table, ii) data was copied across the

production table while assigning a unique id, and iii) the staging table was emptied. This procedure was adopted because in reality the imported data usually needs to be processed and assigned a unique identity in order to link to other tables. PostgreSQL performed >5 times better than MySQL, PostgreSQL took ~445 seconds compared to ~2,940 seconds in MySQL-InnoDB and 2,460 second in MySQL-MyISAM. The actual import script was published as a Supplementary File with my benchmarking article (Khushi 2015).

Data upload performance is critical for bioinformatics servers where many users insert a large amount of data at once. This benchmark identified that PostgreSQL inserts and imports data much faster than MySQL.

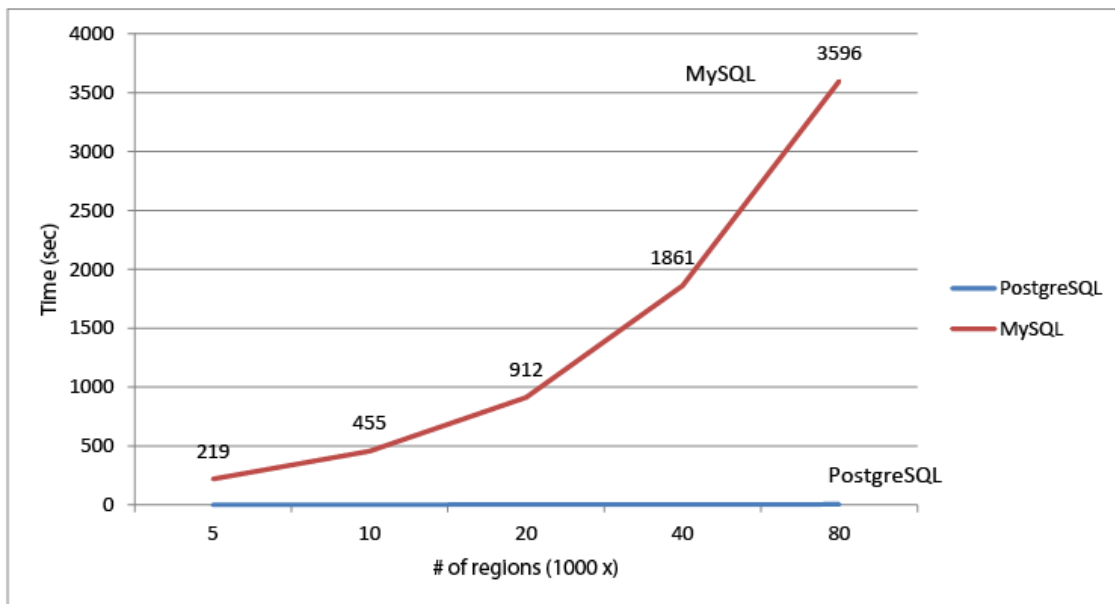


Figure 2-3: Comparison of region insertion performance. For simplicity, MySQL times shown are for InnoDB storage engine as MyISAM did not perform as well.

2.8.2. Benchmarking for Identification of Overlapping Regions

I further investigated the performance of reporting intersecting or overlapping regions using RegMap and using the database built-in functions in each database. The two databases have built-in functions that can be used to identify intersecting lines. Since these built-in functions

are usually used in geographical (spatial) mapping software I subsequently refer to the built-in functions as Geo functions.

PostgreSQL’s performance was again outstanding in finding overlapping genomic regions compared to MySQL (Figure 2.3). RegMap in PostgreSQL took 134 seconds to report intersecting regions for two datasets of 80K regions each, and 257 seconds using PostgreSQL-Geo function. MySQL performance was tested using InnoDB and MyISAM storage engines. MySQL-MyISAM performed poorly compared to InnoDB, however, both engines demonstrated inferior performance as compared to PostgreSQL. For example, when two datasets of 80K regions were tested for overlaps using RegMAP, MySQL-InnoDB took 1,119 seconds, and MySQL-MyISAM took 1,150 seconds. Therefore, for simplicity reasons, I presented the data for MySQL-InnoDB in Figure 2.3.

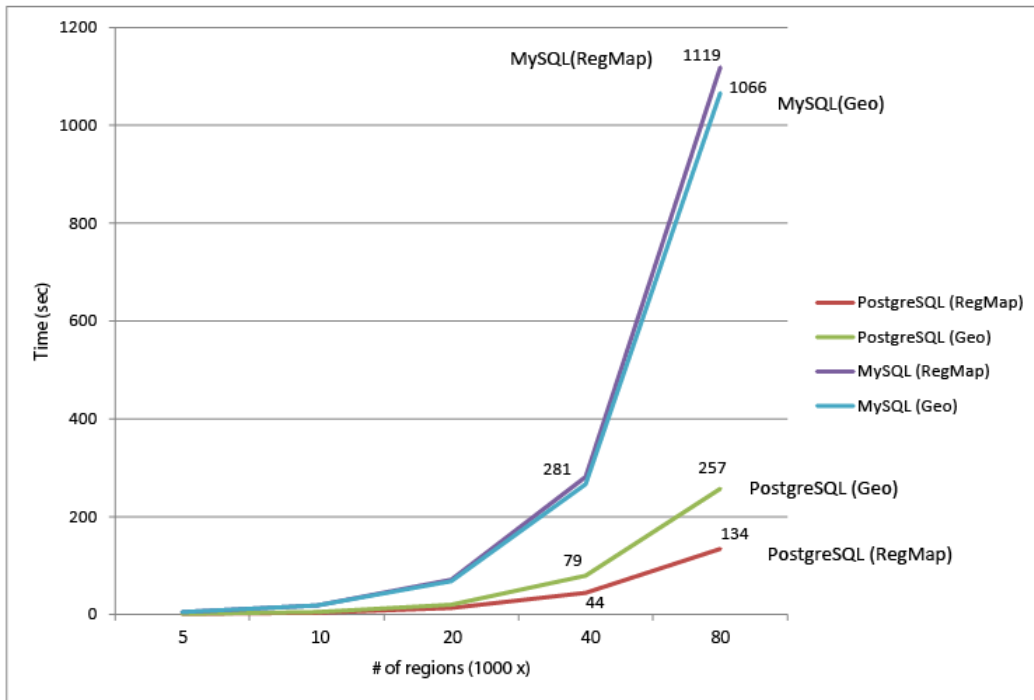


Figure 2-4: Comparison of performance for identifying overlapping regions using RegMap and Geo functions. For simplicity, InnoDB times are shown for MySQL, as the MyISAM storage engine in MySQL did not perform as well.

Primary key fields by default generates clustered index, in addition, performance was tested by setting non-clustered index on various fields and PostgreSQL various index-types: B-tree, R-Tree, GiST, GIN, and Hash. Applying these additional indexes did not improve performance in PostgreSQL while it had a negative effect in MySQL for both engines (InnoDB & MyISAM). I performed these tests on different computers and obtained slightly different timings, however, the overall outcome remained the same which was that PostgreSQL performance in identifying overlapping regions was much better than MySQL. Since RegMap identifies overlapping regions by calculating ‘bp overlap’ for each region, I finally concluded that queries that require extensive calculation of mathematical operations perform much better in PostgreSQL.

2.8.3. Searching and retrieving regions

PostgreSQL was slightly better at performing a search of genomic regions than MySQL. The knowledge base of ~24 million genomic regions was searched for erroneous regions with a start coordinate less than 0 or an end coordinate less than start. PostgreSQL identified 10 erroneous regions in ~5 seconds while MySQL-InnoDB found the same erroneous regions in ~21 seconds and MySQL-MyISAM in 6 seconds. Implementing various types of indexes on chromosome start and end fields did not improve performance for this query. However, searching for specific regions within a certain distance of a gene was instant in all databases. For example, searching regions within 100,000 upstream/downstream of the transcription start site of MYC gene (chr8:128748314) returned results in 3-5 seconds in all databases which was further reduced to ~1 second by implementing an index. Therefore I concluded that the general searching capability of PostgreSQL and MySQL is similar.

2.8.4. Advantages of RegMap over Geo functions

The RegMap algorithm generally performed better than Geo functions, in addition, it provides extended functionality that the Geo functions do not provide. For example, Geo

functions only return a Boolean (true/false) value if the queried regions intersect or not. On the other hand, RegMap reports the number of bases common in the two regions or away from each other. Therefore it is easy to limit results based on the minimum number of bases that must overlap. It also provides the ability to restrict results based on the distance from the centre of regions; this is useful in returning regions that do not share common bases, but are present in close proximity within a specified distance. For example the SQL query *select * from vwregions where bpooverlap<1 and centredistance<1000;* will return regions that do not overlap however their centres are within a distance of 1000 bases. This type of analysis is important in identifying partner factors that bind on DNA in close proximity to each other without overlapping.

I performed benchmarking natively using each product's query interface, however, in the real world the database would be queried using front-end software and it is required that database and designed software have efficient way of handling data uploads. With all databases, a single SQL 'INSERT' statement is used to insert a single row into a database table. However, if a large number of rows are to be inserted, this is a very slow process due to additional computational tasks that are required to run with each transaction. To address the challenge, the databases include special statements to bulk import large number of rows such as SQL Server has BULK INSERT, MySQL has BULK COPY and PostgreSQL has COPY command. However these commands cannot pre-process the data and the format of the imported data needs to exactly match with the table structure where data needs to be inserted. This issue is usually addressed by importing data into a staging table and then after processing the data is transported into the target table. However genomic region files produced by different peak-callers have different numbers of columns (Wilbanks and Facciotti, 2010) creating a difficult scenerio for efficient importing into a database using bulk insert statements.

Microsoft SQL Server provides another unique way of bulk inserting using ‘table-valued parameters (TVP)’. The table-valued parameters, released since Microsoft SQL Server 2008 R2, are special types of variables that behave like temporary tables and all genomic values can be pre-processed, populated into a TVP and passed to the database engine in a single step. This enables inclusion of a complex business logic in a single routine and reduces round trips to the server. In my testing using table-valued parameters Microsoft SQL Server outperformed the other two databases for pre-processing and inserting data into a database table using TVP.

SQL Server being a Microsoft product, the installation and support is better on Microsoft Windows. Therefore I decided to use Microsoft SQL Server for the development of BiSA for Windows however as there was no SQL Server version for Linux/Unix therefore I chose PostgreSQL over MySQL because of better performance.

2.8.5. Other Considerations for Choosing Between MySQL or PostgreSQL

In above benchmarking PostgreSQL clearly outperformed MySQL, moreover, PostgreSQL conforms to International Standards Organization (ISO) standards for SQL, while MySQL does not conform to ISO standards. Transactions in PostgreSQL are more reliable and a programmer always gets the same results without faults (Conrad, 2006). PostgreSQL is released under the PostgreSQL License which is an open source license similar to the BSD or MIT licenses and is heavily used in the industry. Furthermore, the Galaxy development team prefers PostgreSQL because it works better with the SQLAlchemy (Copeland, 2010) database abstraction layer (<http://wiki.galaxyproject.org>). Therefore, I decided to use PostgreSQL for developing BiSA for Unix-like operating systems. Figure 2.4 provides an overview of the process that I used in selection of a database.

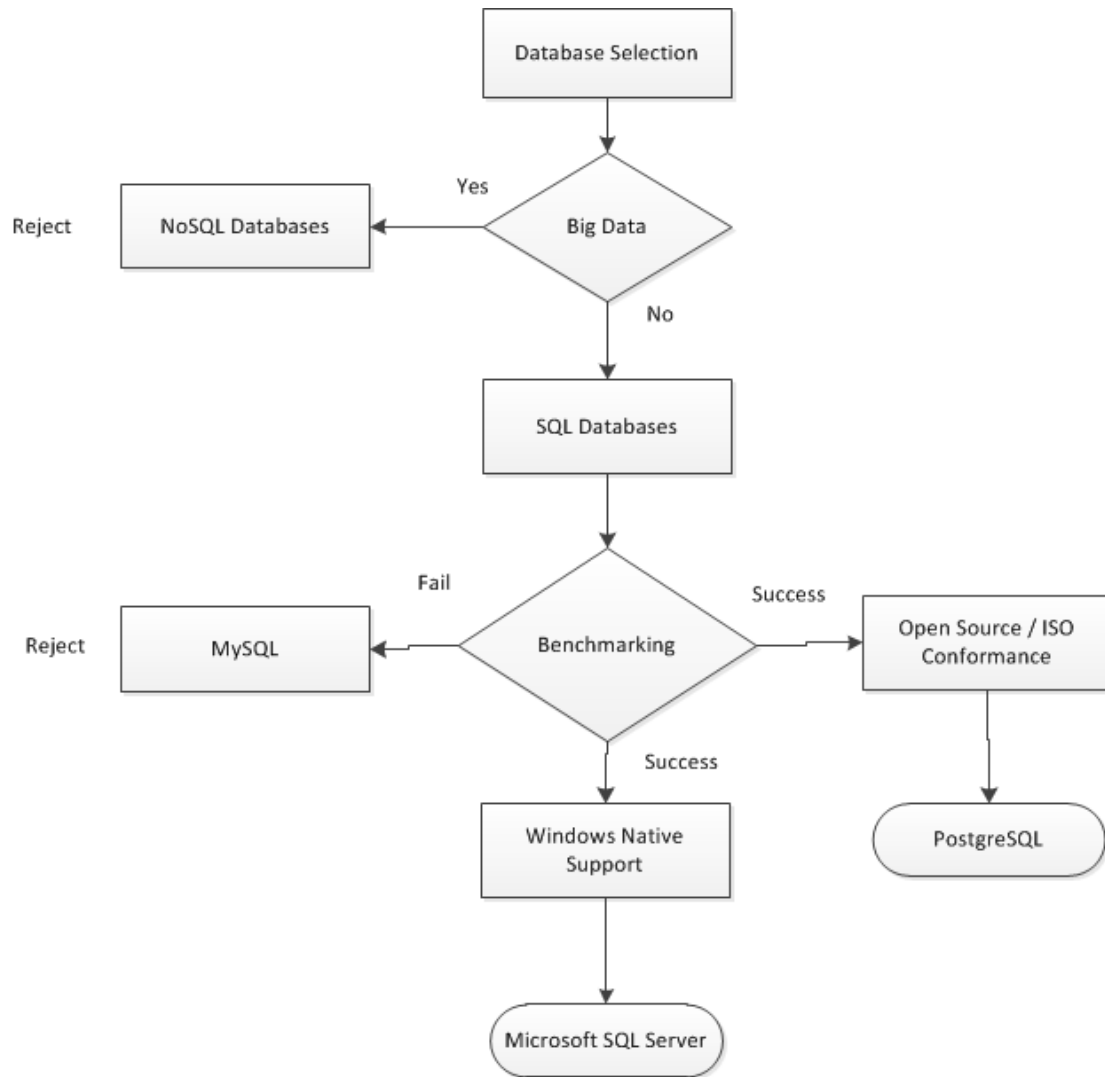








Figure 2-5: Flow chart describing our decision for the selection of a database. Oval represents final successful choices.

2.9. Language Selection for Writing BiSA

After deciding on the database the next step was to decide on a computer language for the development of front-end of the BiSA. Table 2.3 provides an overview of the operating system support, complexity, scripting, object orientation and whether the code run is compiled or interpreted for seven languages. A native compiled code is an executable programme that can be run by computer without any additional software aid, however, interpreted code needs an interpreter software whenever it is run. The native compiled code

runs faster (Sites et al., 1993). On the other hand, a scripting language is usually high level language which is interpreted by another programme at run-time and it is usually easy to write programme in a scripting language (Ousterhout, 1998).

Languages	Operating Systems	Complexity	Scripting	Object Oriented	Compiled / Interpreted
C/C++		+++++	✗	✓	Compiled
Python		++	✓	✓	Interpreted
Java		++++	✗	✓	Interpreted
C# / VB.Net		++	✗	✓	Compiled
Perl		++	✓	✓	Interpreted
Ruby		++	✓	✓	Interpreted




 Windows  Linux/Unix  Apple Macintosh

Table 2-3: Overview of seven main computer languages.

Historically C and its object-oriented extension C++ were designed to write a new operating system (Unix), therefore, it is best for writing low-level applications such as hardware drivers, compilers and system tools and in this function there is no competitor to C/C++.

However the code is complex for trivial programming tasks and database interaction.

Java is a popular language for writing cross-platform GUI (Graphical User Interface) applications that run on different operating systems and devices such as phones, tablets and

computers. The Java community has also developed BioJava for computational biologists that has many built-in bioinformatics functions such as BLAST parser, sequence manipulation, genetic algorithms and statistical distributions (Holland et al., 2008; Prlić et al., 2012). However, Java code is complex due to native objected-oriented support and enforces writing code to handle every possible error. Java like Python, Perl and Ruby is an interpreted language and converts programmes into its custom byte code before execution. The Java and Python compiled form of byte code, also known as object code, cannot run standalone natively and needs a run-time environment software. Therefore, their code performs slower than code written in languages that compile in native code. Python is often quoted as a rapid application development language and it is argued that a programmer can be much more productive in Python than in other languages (Ferg, 2011).

Visual Basic .NET (VB.Net) and C# are two very different languages based on their syntax, classes and history, however, in the Table 2.3 I only have one row for them because both are Microsoft proprietary languages that are exclusively used to write applications run for the Windows operating systems. Both languages share the same runtime engine, both can access any .Net Framework object and share the same integrated development environment 'Visual Studio'. Generally applications written in C# and VB.Net are not compatible on other operating systems, however, Mono open source project (<http://www.mono-project.com/>) can make the applications written in these languages compatible to Mac and Linux operating systems (Nishimura and Timossi, 2006; Kilgore, 2002).

Ruby programming language is fully object oriented and has more advanced functions than Perl such as lambdas and procedures (Flanagan and Matsumoto, 2008). Lambda in Ruby is an in-line function which is handled as an object. The code is more condensed and readable and the Ruby development team has also developed BioRuby (Goto et al., 2010) which has most

features of other Bio-type languages such as BioJava, BioPerl, BioPython. However, there are not many bioinformatics tools that were written in Ruby and there are limited numbers of developers available. Therefore I did not consider Ruby any further.

Perl is frequently used in bioinformatics, due to its heavy use during the human genome project in the 1990s (Wall et al., 2000). At the time Perl provided much needed strings processing capabilities to process long string of sequences and had a strong community base. Subsequently many bioinformatics tools and scripts such as BioPerl (Stajich et al., 2002) were developed. Perl has strong documentation and many books are written on Perl for Bioinformatics (Jagota, 2004; Tisdall, 2009; Tisdall, 2010). Perl is also the language of choice for the Ensembl genome browser. In research, Python and Perl are relatively traditionally established and consequently more widely used in the field of bioinformatics (Dudley and Butte, 2009; Kinser, 2010; Tisdall, 2009; Model, 2010). However, in contrast to Perl, Python gained much popularity in bioinformatics as a scriptable language that is object oriented from its outset (Stajich and Lapp, 2006b). In addition to the features typically found in other scripting languages, Python is useful for bioinformatics and other research because of its various scientific capabilities; thanks to NumPy, SciPy, Biopython (Cock et al., 2009) and many other open projects (Eric Jones, 2001; Oliphant, 2007) providing a large library of functions that solves many common problems in bioinformatics and other research fields. Python's clear and easy syntax makes it suitable for beginners as well as experienced programmers (de Hoon et al., 2003). One of the distinct features of Python which is not mandatory in other languages is that it enforces indentation of nested blocks relative to each other. This makes understanding and interpretation of code very easy.

Most Unix-like operating systems come with pre-installed Python, and many bioinformatics tools such as Galaxy are completely written in Python. As later explained in the Section 2.17

that we decided to build a web-based version of BiSA that run under Galaxy, therefore, I decided to use Python to write BiSA for Unix-like operating systems. A Python version for Windows is also available, however, it is not natively supported and needs to be installed in addition to the database system which might be too much work for biologists. Therefore I decided to use the C# language to write BiSA for Windows because the final code is compiled to an executable (.EXE) file which can be downloaded and run straight away without any additional installation requirement. Figure 2.5 diagrammatically shows our decision for the selection of a computer language.

2.10. BiSA Database Schema

BiSA employs a rational database management system-based architecture to archive unlimited numbers of binding datasets in a very flexible and convenient format. The BiSA database schema is straightforward. All information about a dataset such as factor label, cell line and condition are saved in the 'kbdetails' table while the genomic region data are saved in the 'kbsites' table and linked to the 'kbdetails' table by an identity (KBId) (Figure 2.6).

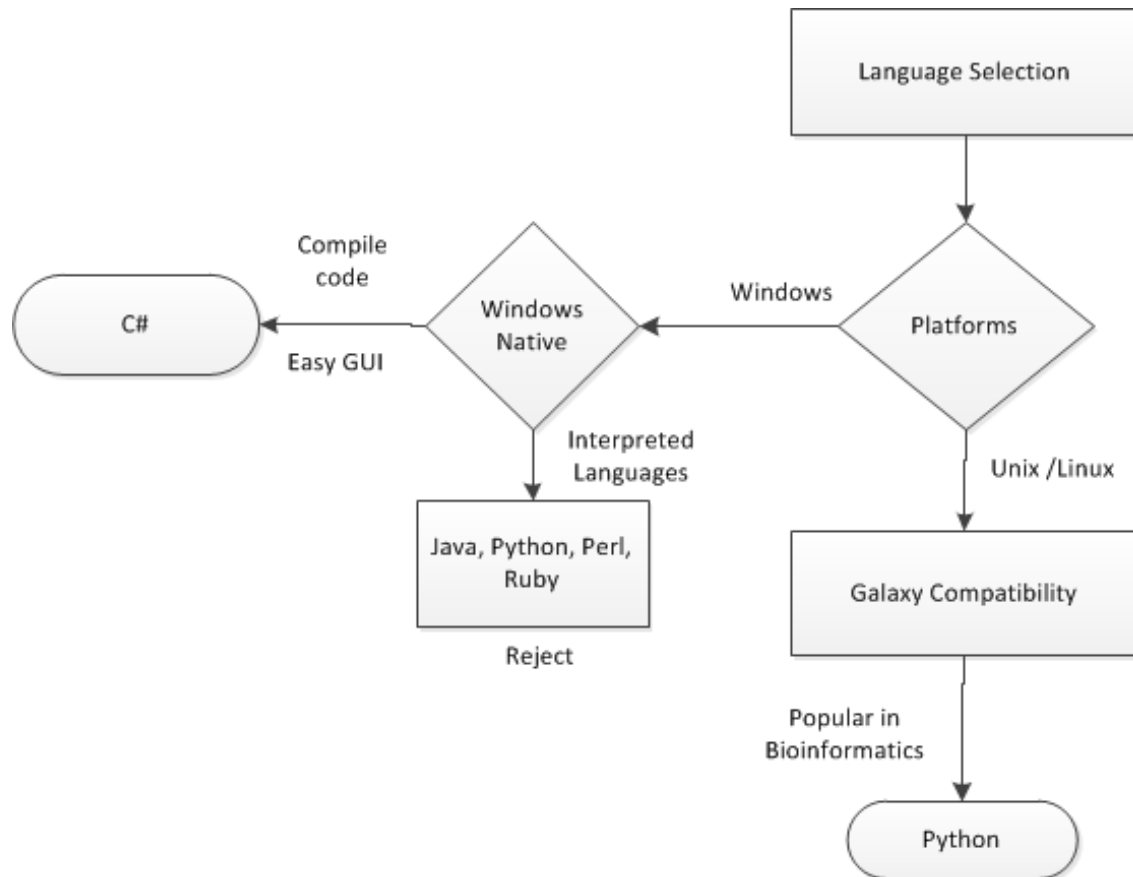


Figure 2-6: Flow chart describing selection for a language for the development of BiSA.

Oval shows final selection of the language.

There are four gene annotation tables for the four genomic assemblies hg19, hg18, mm9 and mm8 genome assemblies having exactly the same fields, therefore shown as one table structure in Figure 2.6. These annotation tables are not linked to KbDetails or KBSites tables. The email field in the KbDetails table is only used in the PostgreSQL database in BiSA for Linux web-based version to track the ownership of the datasets.

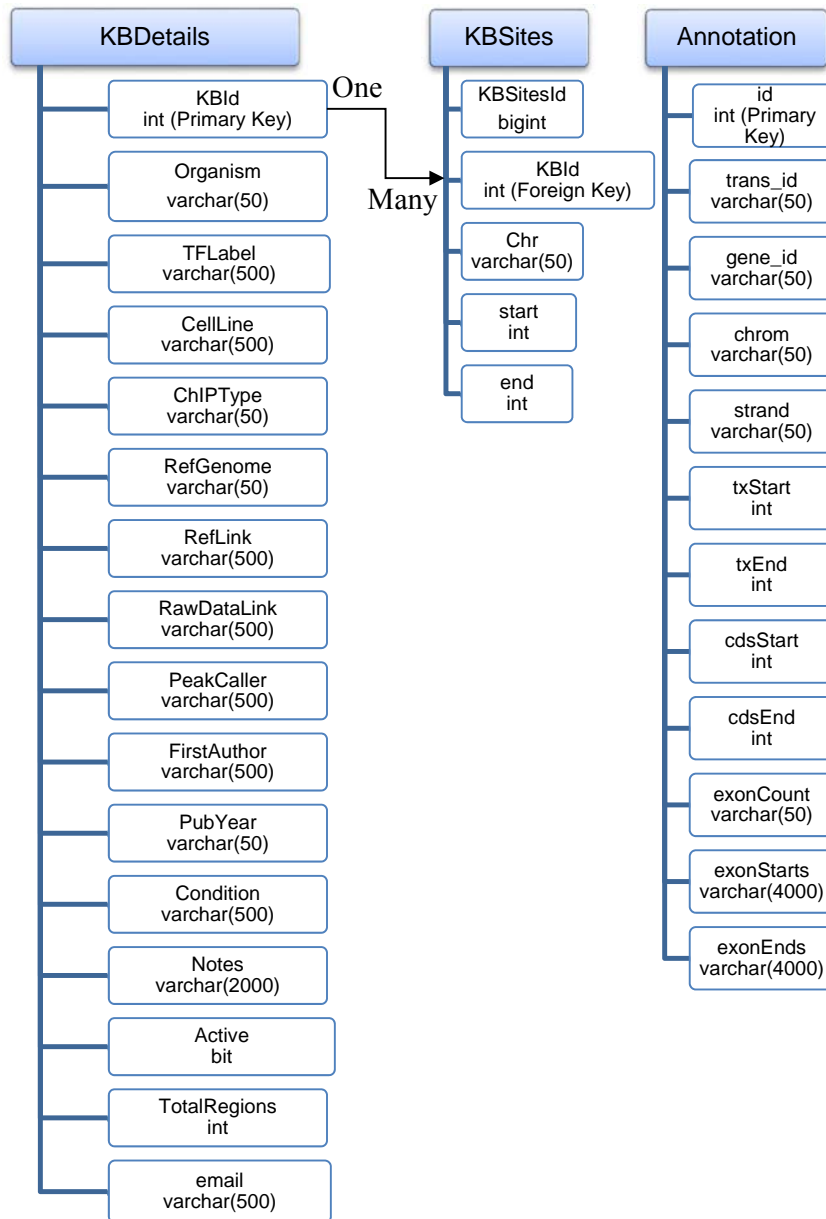


Figure 2-7: BiSA Database Schema. Every region in the KBSites table is linked with the KBDetails table by the KBId (Primary key in KBDetails, foreign key in KBSites table). Clustered indexes were generated on primary keys by default, no other index-type were created. Four annotation tables were created for hg19, hg18, mm9 and mm8 assemblies, having same structure, so shown as one 'Annotation' table. The email field in the KBDetails table is only used in PostgreSQL database in BiSA for Linux version.

I have also used temporary tables in some stored procedures to extract overlapping or non-overlapping regions. The temporary tables are created on run-time in a SQL system database 'tempdb' whose collation could be different from the collation of the BiSA database. A database collation is a set of rules for dealing and comparing characters in character set. The conflict of collation makes matching of characters difficult so I have written code for such stored procedures that run-time check and modify the collation for temporary tables.

Microsoft SQL Server databases generate transaction logs that keep records of every transaction performed in the database. The transaction log could grow very fast depending upon the usage and the type of recovery model in use. Large logs could slow or stop the overall functionality of the database. There are three recovery models i) simple, ii) full and iii) bulk-logged. I have used the simple recovery model to keep the transactions logs small, however, the simple recovery model only enables the retrieval of data from the recent backup of a database (Bernstein et al., 1987; Haerder and Reuter, 1983).

2.11. BiSA Application Architecture

There are three main layered components of the BiSA, i) Graphical User Interface (GUI), ii) the framework and iii) the database system. I have used Microsoft Visual Studio 2010 to design the GUI as the Windows form and middleware application was written using C# and Microsoft .Net Framework while Microsoft SQL Server is being used as database system. For the Linux environment the GUI is web-based written as Galaxy forms, while Python is used to interact between with the PostgreSQL database (Figure 2.7).

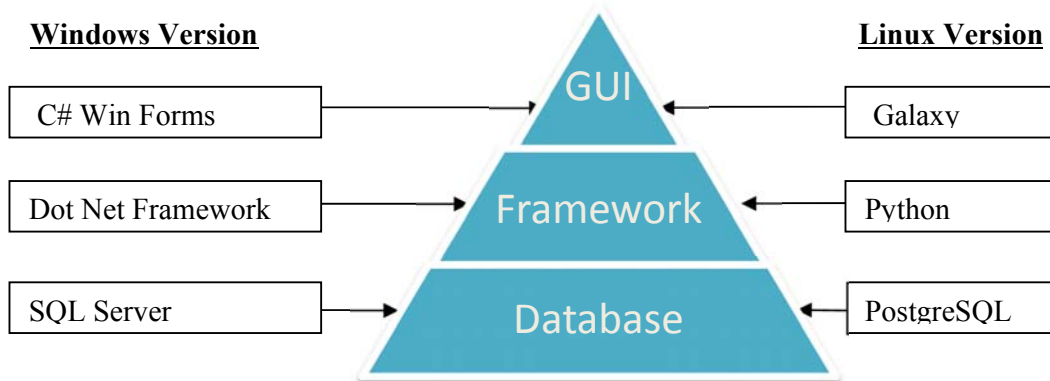


Figure 2-8: BiSA application architecture.

In both Windows and Linux scenarios the database server can be hosted on another server on the network to reduce the work on one computer and to speed up the overall performance.

Figure 2.8 diagrammatically shows the client-server architecture.

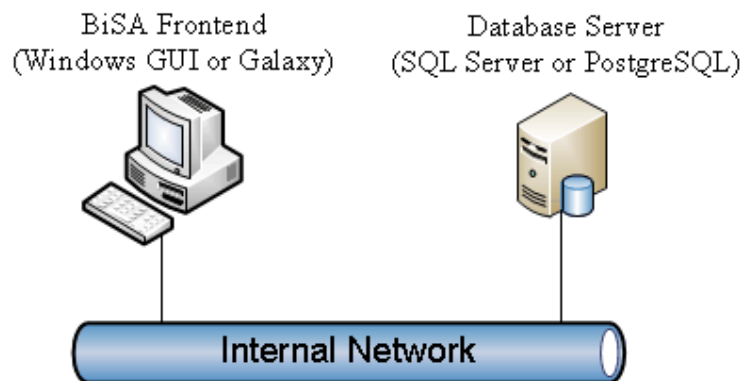


Figure 2-9: BiSA in a client-server architecture.

2.12. BiSA Charts

For generating Venn diagrams and histogram charts, I have written an application programming interface in C# to interact with Google Charts (Charts, 2014). Google Charts can draw a Venn diagram for a maximum of three datasets. In the application programming interface the size of each circle and the size of the overlap between the two datasets are specified. If there are three datasets to be compared then a number representative of the overlap of the three datasets is also required.

Initially, I devised two algorithms to report overlapping regions among three datasets, however only the stringent algorithm that reports lesser overlapping regions is implemented in BiSA. To explain the algorithms let us consider three datasets. In the first algorithm, the overlapping regions of the two datasets are extracted and then these overlapping regions are checked for overlapping regions in the third datasets. For example, in Figure 2.9-A region A overlaps with two other regions B and C. Therefore region A is said to overlap the other two regions. When the regions of the first datasets that overlap with the other datasets are required to extract then this algorithm is implemented. However for the drawing of Venn diagrams we use the algorithm explained below.

In the second algorithm the two datasets are checked for overlapping regions. The overlapping sections of the overlapping regions are extracted and checked for the overlapping regions in the third dataset. For example in Figure 2.9-B the overlapping section D is extracted from the overlapping regions A and B. This overlapping section is then overlapped with the third region C. This algorithm is more stringent and requires at least 1 base pair in common in the three datasets. For drawing of the Venn diagram we calculate the number of common regions using this algorithm.

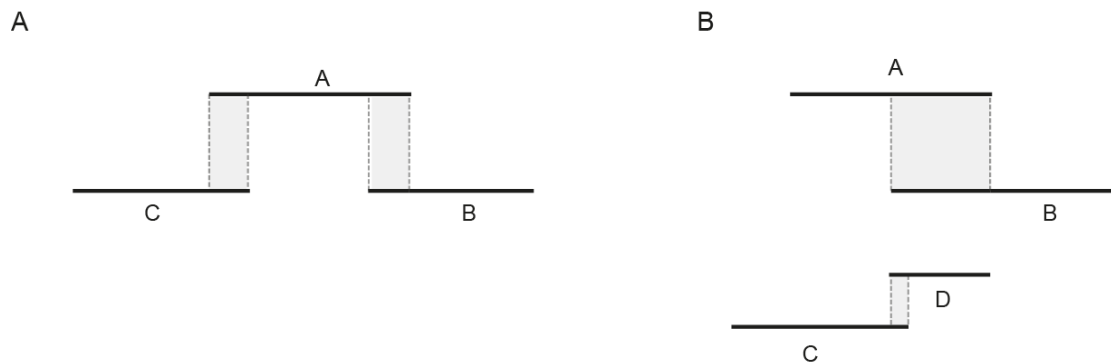


Figure 2-10: Identifying three overlapping regions. A) Region A overlaps with regions B and C. B) The overlapping section D of the overlapping regions A and B is extracted and if the

section D overlaps with region C then the three regions A, B and C are reported as overlapping regions. We have implemented this stringent algorithm in drawing Venn diagrams to show the extent of overlap among three datasets.

2.13. Statistical Significance

We have implemented IntervalStats (Chikina and Troyanskaya, 2012) in BiSA to test the statistical significance of overlap between two dataset. IntervalStats is a command line tool written mainly for the Unix environment. Therefore, we used the MinGW toolkit (MinGW, 2013) to compile it for the Windows environment. The BiSA for Windows download package includes an IntervalStats executable file and dependent Dynamic Link Library code files, however, the tool runs independently of BiSA. When the IntervalStats tool is executed through the BiSA GUI, the datasets under study are saved in the ‘data’ subfolder. During the execution of the statistical tool the terminal window stays open to show the messages from the tool. IntervalStats calculates a p-value for each region in a query dataset against the nearest region from a reference dataset. A defined domain dataset, representing the line-space of all possible interval locations, acts as a background to the statistical test and can be restricted to specific locations, such as promoter proximal regions, to take into account known biases in binding site detection. In the simplest case, the domain comprises the entire genome. We have populated BiSA with a number of domain files such as promoter regions within 10kb of a TSS, intergenic regions and whole genome for hg19, hg18, mm9 and mm8 assemblies. Users can select one of the prepopulated domains or can specify a BED file as the domain. In addition to individual p-values for region overlap, IntervalStats returns a summary statistic, referred to as the Overlap Correlation Value, to identify the overall significance of overlap of two datasets. This summary statistic represents the fraction of regions in the query dataset with a p-value of overlap to the reference below a significance threshold value, and thus reflects the likely significance of overlap of the query and reference

datasets. The correlation coefficient can range from 0 to 1, the closer the value to 1 the stronger the significance of overlap of two datasets. We have set the threshold p-value to 0.05, however this value can be changed in the configuration file, BiSA.exe.config if desired.

2.14. Gene Annotations

BiSA also provides the functionality of annotating the regions of interest with known genomic features. For this functionality gene annotation data are obtained from the UCSC genome browser. Initially we have populated annotations for reference genomes hg18, hg19, mm8 and mm9. Custom gene definitions or additional genomes for other organisms can be uploaded in the software by the user.

2.15. BiSA on Sourceforge.net

Since many bioinformatics tools are available free of charge and the community welcomes any new open source tools, we released BiSA as an open source free software available under GNU General Public License. Initially we hosted BiSA code base on CodePlex, Microsoft's free open source project hosting site because of better compatibility and options for Microsoft .Net Framework projects. However when I started developing BiSA for Linux using Python language and PostgreSQL database, we decide to re-locate BiSA on Sourceforge.net. The link for the project website is <http://bisa.sorceforge.net>. I also prepared download packages for the Windows and Linux environment and tested the download and installations on different computers and platforms.

2.16. BiSA for Windows: Installation and Configuration Testing

BiSA for Windows employs Microsoft SQL Server, therefore, I tested BiSA on various recent versions of SQL Server including SQL Server Express versions. There are three main steps of installation of BiSA for Windows:

- i) Installation of Microsoft SQL Server database engine

- ii) Downloading and restoring the BiSA database file
- iii) Linking of the front-end application to the database

The installation of SQL Server could be a non-trivial task for many biologists therefore before publishing BiSA the installation, restoring and linking of BiSA were tested on various versions as explained below:

i) Installation of Microsoft SQL Server database engine:

There are many editions of SQL Server such as Express, Standard, Enterprise, Developer etc. I developed BiSA to work on all of these SQL Server database platforms. The Express Edition is limited to 1GB memory utilization and maximum database size is 10GB. Due to this limitation BiSA performance might be little slower on the Express Edition, therefore, it is recommended that a non-express edition is employed such as SQL Server 2008 R2/2012/2014 Developer or Standard Edition. A free copy either from DreamSpark (Microsoft, 2013a) or from WebsiteSpark (Microsoft, 2013c) can be obtained. Most educational institution would have a licensed version under Microsoft Enterprise Agreement. Otherwise the free Express edition is available at <http://www.microsoft.com/sqlserver/en/us/editions/express.aspx>.

I downloaded various versions of the databases along with the Microsoft SQL Server Management Studio (SSMS). SSMS is a visual tool to manage SQL Server which can be downloaded separately for SQL Server 2008 R2 from Microsoft download centre (Microsoft, 2013b). The Microsoft SQL Server 2012/2014 Express edition is supplied with SSMS. The Microsoft SQL Server database engine installation wizard prompts to assign an instance name to the new installation. I named it 'sqlexpress' however a different name can be assigned, but the name must match the name in the ConnectionString of the BiSA

configuration file (BiSA.exe.config) explained in forthcoming sections. After completing the installation of SQL database engine and the SSMS, SSMS can be used to restore the BiSA database (download from the project website <http://bisa.sourceforge.net>) as explained below.

ii) Downloading and restoring the BiSA database

The BiSA download includes the backup database file named BiSAx.xx.bak. To restore the file, Microsoft SQL Server Management Studio (SSMS) is opened and connected to the database engine. Right clicking on the ‘Databases‘ folder icon under Object Explorer and choosing ‘Restore Database...’ brings up the Restore Database screen (Figure 2.10).

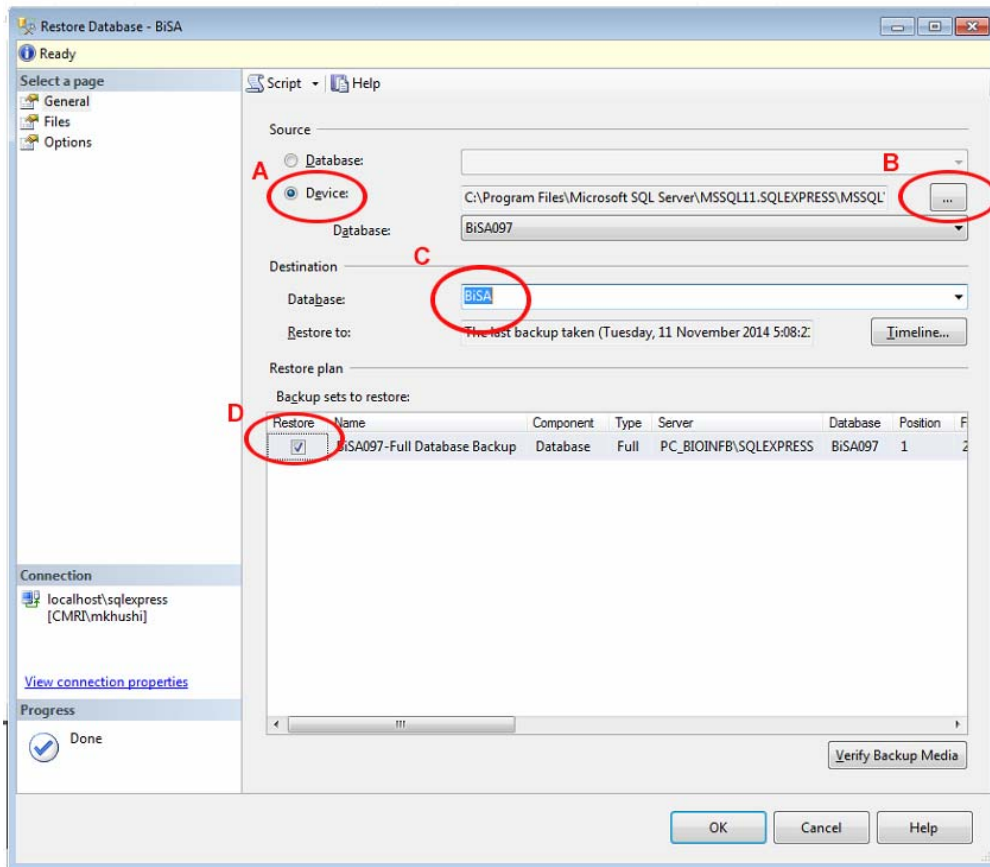


Figure 2-11: Restore database screen. A) Clicking ‘Device’ option makes “...” button (B) available for the selection of backup set. Specify a name of restore (C) and click the available backup set (D)

Finally clicking OK (Figure 2.10) restores the database to its original folder location of the backup where the backup was taken from. If the target system does not have the expected folder structure then Microsoft SQL Server gives the following error message:

"System.Data.SqlClient.SqlError: Directory lookup for the file "C:\path\BiSA.mdf" failed with the operating system error 21(The device is not ready.). (Microsoft.SqlServer.Smo)"

The BiSA backup was created on Microsoft SQL Server 2008 R2 therefore this problem can also happen if the restore is attempted on Microsoft SQL Server 2012/2014 that has a different folder structure. In that case the database has to be restored in another folder. Different folder than the default can be specified by choosing “Relocate all files to folder” under Files option on the Restore Database screen (Figure 2.10).

Detailed steps with screenshots for restoration of BiSA database are described on the project website (<http://bisa/sourceforge.net>).

iii) Linking of the front-end application to the database

The BiSA download package also includes an XML-based configuration file named ‘BiSA.exe.config’. The configuration file holds the variables where the name of the database engine, database name and other settings are defined. More about configuring this file is discussed in section 2.15.

2.17. BiSA for Linux/Galaxy

Unix, Linux and Mac share a similar operating system environment and with some effort applications can be written to support all of these platforms. Therefore BiSA is written to be run on any of the Unix-like operating systems. However, instead of writing our own web-based front end, we decided to integrate BiSA with the Galaxy (Novak et al., 2013;

Blankenberg et al., 2010; Giardine et al., 2005) genomic research web-based platform for the following reasons:

- i. Galaxy is an open source platform and provides a user-friendly web interface (Goecks et al., 2010b) (Schatz, 2010).
- ii. Users can download and install Galaxy locally and any tool that runs on Unix/Linux command line can be integrated in Galaxy (Taylor et al., 2007).
- iii. Many recently developed bioinformatics tools are Galaxy friendly (Liu et al., 2011b; Barash et al., 2013; Cock et al., 2013).
- iv. Galaxy provides eXtended Markup Language (XML) based programming platform which makes it easy to integrate existing tools to the Galaxy framework.
- v. Galaxy has a large user base and extensive documentation is available.
- vi. All tools run on a web server and a client does not need to install any tool.
- vii. Galaxy has been utilised by many Australian universities and has been used in the GVL (Genomics Virtual Lab) project which is funded by NeCTAR, an Australian Government project.

2.18. Tool Integration in Galaxy

Galaxy provides a XML based tool configuration script (GalaxyTeam, 2013) which outlines a tool's inputs as web form fields such as text boxes, dropdown option and checkboxes and also defines and links the output of the tool to the web interface. XML is a mark-up language like HTML (Hyper Text Markup Language) defined by W3C (World Wide Web Consortium) (Bray et al., 1997), however unlike HTML, tags in the XML are not defined. Any text can be defined as a tag following some simple rules. Galaxy has defined its own XML tags to perform various options such as *command* tag specifies the language of the tool being used and all the parameters that will be passed to the tool. The Box 2.3 is a code extract of Galaxy XML that I wrote to import datasets from Galaxy into the BiSA database, the code also

checks the data type of the imported dataset if it is a GFF or BED dataset and parses parameters accordingly.

```
<command interpreter="python">importdatasets.py $__user_email__

    $input  $organism  $refgenome $cellline $dataLabel

    #if isinstance( $input.datatype,
    $__app__.datatypes_registry.get_datatype_by_extension('gff').__
    class__):
        gff

    #else:
        bed

    #end if

    $out_file1

</command>
```

Box 2.3: An example of *Galaxy XML code*. The *command* tag specifies the language of the tool and sequence of the parameters that will be processed by *BiSA import command importdatasets.py*.

Since BiSA for Galaxy is a web-based tool and intended to be used by many researchers at the same time, we also save the email address of the researcher when a dataset is imported as explained in Section 2.11 BiSA Database Schema. This ensures that only the owner of a dataset is allowed to delete or analyse its data. Galaxy's built-in variable `$__user_email__` provides the email address of the logged-in user, which is passed to BiSA import dataset command (Box 2.3).

2.19. BiSA for Galaxy: Installation and Configuration Testing

I developed BiSA to be compatible to run on all variants of Unix-like operating systems therefore I tested the installation of BiSA on various operating systems such as Red Hat Enterprise Linux (RHEL), Ubuntu and Mac. At the University of Sydney the RHEL server was hosted behind a firewall that required authentication, therefore, the following shell command was used to set the proxy.

```
export http_proxy=http://username:password@web-cache.usyd.edu.au:8080
```

Setting this proxy was also important to obtain Galaxy software updates automatically as whenever Galaxy was run, it checked for updates by default. By default Galaxy uses SQLite database, however, we have also chosen PostgreSQL database to write BiSA as explained above. PostgreSQL was downloaded from *www.postgresql.org* for various platforms.

I installed PostgreSQL on the same server as that of Galaxy, however, if PostgreSQL database server is on a different machine then it is required that configuration file (pg_hba.conf) be configured to allow connections from the Galaxy machine. This file (pg_hba.conf) ensures that the database server is open for incoming connections to PostgreSQL port (default is 5432), in case PostgreSQL is installed on Windows 7 machine then incoming connections to the port 5432 can be allowed by using *Windows Firewall with Advanced Security* under Control Panel.

Once Galaxy and PostgreSQL were installed; BiSA was installed and configured using the following steps.

1. BiSA Linux package was downloaded from the project website (<http://bisa.sourceforge.net>) (source code and the database backup file).
2. A new database named *bisa* was created.

3. The database was restored by the following command:

```
pg_restore --host=localhost --port=5432 --username=your_db_username --password  
--dbname=bisa /backup_location/bisa_pg_0.xxx.backup
```

4. A new directory named *bisa* was created in the *tools* directory of the Galaxy installation typically */galaxy-dist/tools/*.
5. All source code files were copied (*.py and *.xml) in */galaxy-dist/tools/bisa/*
6. A new section was added in */galaxy-dist/tools_conf.xml* by writing the following lines:

```
<section name="BiSA" id="BiSA">  
  
<tool file="bisa/importdatasets.xml" />  
  
<tool file="bisa/browsekb.xml" />  
  
<tool file="bisa/analysis.xml" />  
  
<tool file="bisa/statsign.xml" />  
  
<tool file="bisa/annotation.xml" />  
  
<tool file="bisa/proximalfeatures.xml" />  
  
</section>
```

7. Galaxy was stopped (if already running) and started by the *sh run.sh* command
8. BiSA appeared as a new section on the left *Tools* panel.

2.20. BiSA Developmental Issues

During the development of BiSA I came across many developmental issues, the most important of which are discussed in the following sections.

Some issues stop a programmer from progressing any further. For example, Visual Studio (the GUI integrated environment to write C# applications) crashed and stopped working with the message “ContextSwitchDeadlock was detected”. This error occurs because some of the genomic region processing code took longer to run than normal Windows programming. This

error was corrected by un-checking the ContextSwitchDeadlock option available in the Debug->Exceptions screen and expanding the "Managed Debugging Assistants" options. The ContextSwitchDeadlock option ensures that a programmer of a software interacts with the GUI all the time, however, in the case of genomic region processing this was not required. Another error does not allow the SQL Server to start and attempt to start the database engine gives the error message "*Failed to generate a user instance of SQL Server due to a failure in starting the process for the user instance. The connection will be closed.*" This error was corrected by assigning the Local System account to start up the Microsoft SQL Server service.

Some issues occur all the time and need to be addressed in a better way, such as connecting to the database. Therefore, I have created a separate text-based XML file 'BiSA.exe.config' and saved the connection string in the file. The connection string is a text-based instruction to the application informing where to locate the database server and what username/password should be used to connect the database. An example of the connection string that is used to connect the database on a local computer is below.

```
<add name="BiSA.Properties.Settings.BiSAConnectionString"
  connectionString="Data Source=localhost\sqlexpress; Initial Catalog=BiSA;
  Integrated Security=True" providerName="System.Data.SqlClient" />
```

This connection string *Integrated Security=True* directs the application to use Windows Integrated Security therefore username and password are not required to connect to the Microsoft SQL Server as long as the database accepts incoming connections. The *Data Source* property sets the location of Microsoft SQL Server and its instance name. If SQL Server is installed as a default instance then specifying just the machine name for the *Data Source* is sufficient. The *Initial Catalog* property sets the name of the database. If the SQL

Server is installed on a networked server (Figure 2.8), then it is required to set a database username and password. In this case the following connection string should be used.

```
<add name="BiSA.Properties.Settings.BiSAConnectionString"
connectionString="Data Source=Server_name_Or_Ip_address\instance_name;
Initial Catalog=database_name; PWD=password;UID=username; persist security
info=True" providerName="System.Data.SqlClient" />
```

I have written code to wrap potential errors and display alerts explaining the problem. For example, BiSA displays the following error message if there is a problem in connecting to the database.

“Cannot open database BiSA requested by the login. The login failed. Login failed for user ‘ComputerName\UserName’”.

This error means that there is a problem with the connection string as explained above. The BiSA.exe.config file also holds the value of the threshold p-value to calculate the correlation coefficient as explained in the statistical significance section.

BiSA for Galaxy is written in Python and PostgreSQL database. If PostgreSQL is setup on another networked machine then it is required that the Galaxy web-front machine is authenticated in the PostgreSQL configuration file “pg_hba.conf” (Massa and Riggs, 2013).

Otherwise the following error occurs:

OperationalError: (OperationalError) FATAL: no pg_hba.conf entry for host

2.21. Discussion

In this chapter I described how I have selected various computational tools to write BiSA bioinformatics resource and tools to analyse genomic data. Firstly, I developed a BiSA for Windows version as a desktop application to address needs of an investigator. The Windows

desktop version was written in C# using Visual Studio and Microsoft SQL Server was used as the backend database. Desktop version requires installation of Microsoft SQL Server which could be difficult for biologist, moreover, we identified its limitation with the scalability of useability for large bioinformatics facilities that run their computer systems in Unix-like operating systems. Therefore, in second round of development I built a web-based Unix/Linux/Mac version that runs under Galaxy. The Galaxy version is written in Python utilising PostgreSQL as backend database. To systematically decide on a suitable database for BiSA, I performed database performance benchmarking by writing a Region Mapping (RegMap) algorithm for various databases.

Various databases are heavily used in genomic research, therefore researchers would benefit from knowing which database product performs better for a specific type of data.

Benchmarking software products helps vendors to improve their products and helps users to select a product suited to their needs. Various benchmarks for database systems exist and it is acknowledged that development and adoption of benchmarks advance research in a research area (Sim et al., 2003; Ray et al., 2011). Ray et. al. (Ray et al., 2011) benchmarked databases for spatial data, Xu et. al. (Xu and Guting, 2012) benchmarked database for moving objects data. Similarly various other benchmarking efforts and their benefits are acknowledged (Bose et al., 2009; Venema et al., 2013; Arslan and Yilmazel, 2008) (Aniba et al., 2010). However there is no benchmarking effort exists for database performance on genomic region operations. Therefore RegMap, being natively written in SQL for Microsoft SQL Server, MySQL and PostgreSQL, will advance research in this field and will provide a baseline mark for future algorithms.

ChIP-Seq analyses produce a large number of variant files. Usually detailed information about factor, cell-line, condition, peak-calling or analysis parameters used are recorded as part of file names or kept separate which makes it difficult to manage such information for a

large scale study. Databases provide a more effective way of managing curation, annotation, sorting and relationships among data. Therefore RegMap, being a SQL based algorithm, can be integrated in any language as most languages provide an application programming interface to connect to SQL-based databases. SQL's simple syntax is also easy for a biologist to learn.

There are a number of tools that are in use by the research community to operate on genomic regions, for example BEDTools (Quinlan and Hall, 2010b), Pybedtools (Dale et al., 2011b), GenomicTools (Tsirigos et al., 2012), and BEDOPS Tools (Neph et al., 2012). All of these tools are designed to operate on text files and integration of these tools in other languages is usually difficult. Tabix (Li, 2011) is another efficient tool that is usually used to extract specific regions from large files. However, there is no algorithm available that performs genomic region operations natively in a relational database system. Therefore direct comparison of the performance between RegMap algorithm and other tools that work on files is not appropriate.

RegMap benchmarking identified that PostgreSQL extracts overlapping regions much faster than MySQL. Insertion and data uploads in PostgreSQL were also better, although general searching capability of both databases were almost equivalent. For example, both databases when searched a table with ~24 million real genomic regions, returned results in ~1 second for regions that were within 100K of transcription start site of MYC gene. However we identified that applying database indexes do not improve the performance for these kinds of genomic operations.

RegMap benchmarking shows that there is great deal of opportunity to improve the database built-in functions that are used to find intersecting geometrical shapes. I acknowledge that in other fields such as geo mapping application it is usually not required to find the thousands of intersecting features. However for genomic data such as histone marks, the genomic regions

can be more than 80 kb. Therefore, with increased use of databases in genomic applications, it is needed that database functions are improved and enhanced. I have proposed the introduction of a new 'genomic region' data-type in all databases (Khushi, 2015).

Chapter 3: Binding Site Analyser (BiSA): database resource, archival and tools to analyse genomic regions

3.1. Introduction

The recent revolution in whole genome census approaches has seen an exponential increase in available data sets describing genomic features, such as transcription factor (TF) binding sites and histone modifications. Recent studies have revealed that there are often overlaps and co-association between transcription factors at binding sites (Gerstein et al., 2012; Ballaré et al., 2013) and identifying relationships, such as overlaps in genomic features, has become a fundamental biological research tool (Meyer et al., 2012). Moreover, the existence of a wealth of published public data sets now provides opportunities for data mining in large databases of archived genomic data.

Existing methods of finding overlaps such as BEDTools, UCSC Table Browser, Homer or Segtor (Heinz et al., 2010; Quinlan and Hall, 2010a; Renaud et al., 2011) are limited in functionality for simultaneous comparison to multiple archived data sets. Moreover, few tools provide a simple interface that can be easily implemented by biologists with limited computing skills.

To address these challenges, we have developed BiSA (Binding Sites Analyser), which allows the investigator to analyse overlapping or non-overlapping regions, to visualise results by Venn diagram, and to identify the genes in the proximity of regions of study. BiSA is controlled through a user-friendly graphical user interface (GUI), installed on a Windows environment or embedded in the Galaxy web-based high throughput genomic analysis tool. Both options maximise the ease of use of this powerful tool for molecular biologists, who may lack the necessary computing skills required to use alternate approaches.

In Chapter 2, I discussed the basis on which various computation resources were selected and have described the backend design of BiSA including computational issues. In this chapter I describe the salient features of BiSA version for Windows and Galaxy.

3.2. BiSA for Windows

The BiSA Windows GUI is split across seven tabs, which flow with the usual analysis steps of studying genomic regions, i) importing datasets, ii) selecting datasets to be compared, iii) performing analysis, iv) studying statistical significance of overlapping regions, v) performing annotation of genomic regions, vi) extracting proximal features of specific genomic regions and finally vii) administrating datasets (Figure 3.1).

3.2.1. Import Datasets to Knowledge Base (KB)

This is an optional step as the user can choose to analyse only data already contained in the KB. The user browses for their dataset, which can be imported in tab-delimited or comma delimited BED or GFF format, assigns a logical name and description for the data, and uploads to the KB (Figure 3.1). The first 20 lines of the data can be displayed for verification. Chromosome position is 0 indexed as in BED format. Comments or header information in the file are reported as failed records in the ‘Report’ section. If no valid data are imported in the first 50 lines, the upload fails and BiSA stops the import process. The user enters information about organism and cell line, TF and conditions, which are saved along with the database record. The genome build for the genomic region coordinates must be entered during this process (Figure 3.1, circled) and the record will be limited in future analyses to comparison with other datasets generated in the same genome and build. Associated data and publication links can also be added at this stage.

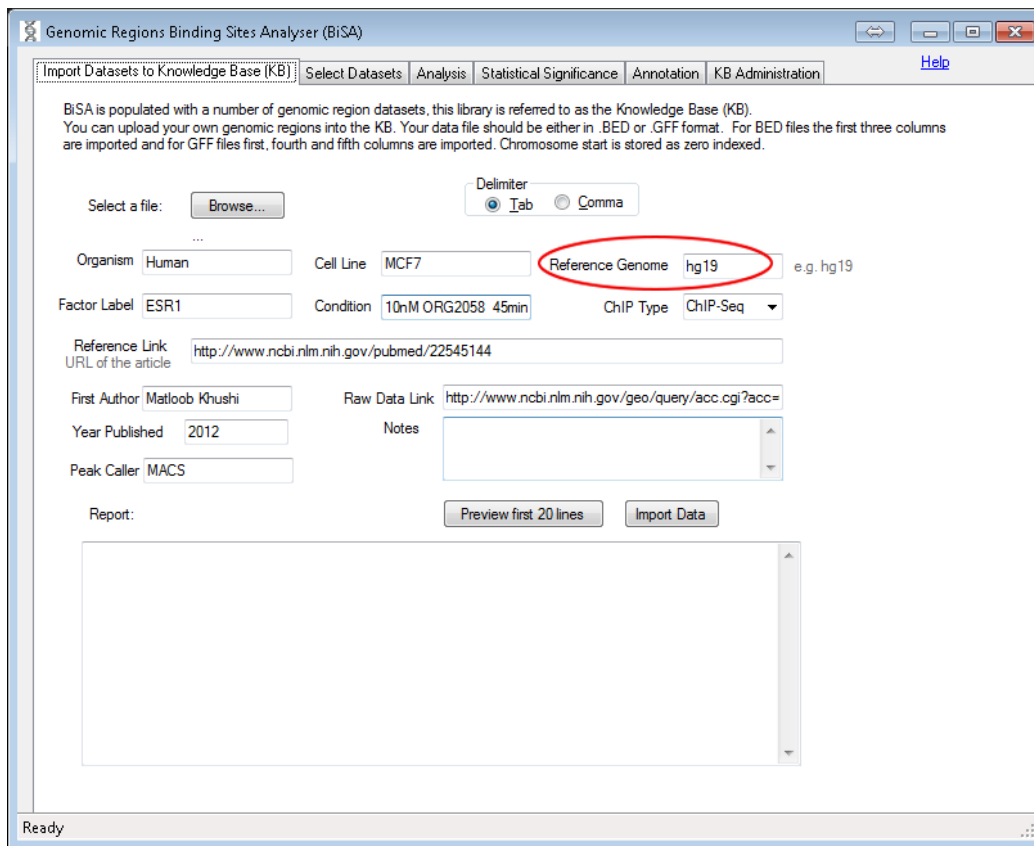


Figure 3-1: Import Datasets to Knowledge Base (KB). This step is optional and users can study data already saved in the KB, without importing datasets. In this step, the user can upload their own transcription factor DNA binding sites or histone modification locations, usually as BED or GFF peak files. If the file extension is other than BED or GFF, BiSA prompts the user to choose the right format. It is important to specify a Reference Genome (encircled), for instance hg18/hg19 for human or mm9/mm8 for mouse. BiSA will only allow comparisons between datasets of the same reference genome.

3.2.2. Select Datasets

This tab displays a list of all datasets in the KB including those uploaded in the first import tab. Data are selected for analysis by checking the "active" box beside the relevant dataset (Figure 3.2A). Only data from matching reference genomes can be selected for analysis. A checked tick in the 'Active' column represents an active dataset that can also be used in the third analysis tab, and only active datasets can be annotated. To change the active status of datasets from one reference genome (e.g. hg18) to another (e.g. mm9), the user must

deactivate all datasets first, which can be done by toggling on and off the ‘Select All’ check box and pressing the Update button. Clicking on the text of any row displays further information about the data. Website addresses are hyperlinked to the source websites/articles for the data. After selecting datasets for analysis, clicking on the Update button activates datasets in the BiSA database. The search field (Figure 3.2B) allows the user to search the KB by organism, cell line, factor label, reference genome or peak caller. Only datasets that are active can be displayed by checking the ‘Active datasets only’ option in the ‘Display Filter’ (Figure 3.2C). Displayed data can be sorted according to any of the database fields by selecting the column heading for the field of choice (Figure 3.2D).

3.2.3. Analysis

This is the main analysis screen where users can analyse active datasets. Six types of analysis are provided: a) calculate percentage overlap of all active datasets, b) extract regions that overlap with all active datasets, c) extract overlapping sections of regions common in all active datasets, d) extract regions that overlap between two selected datasets, e) extract regions that do not overlap with another selected dataset, f) extract overlapping sections of regions common in two datasets. Analysis can be restricted by chromosome.

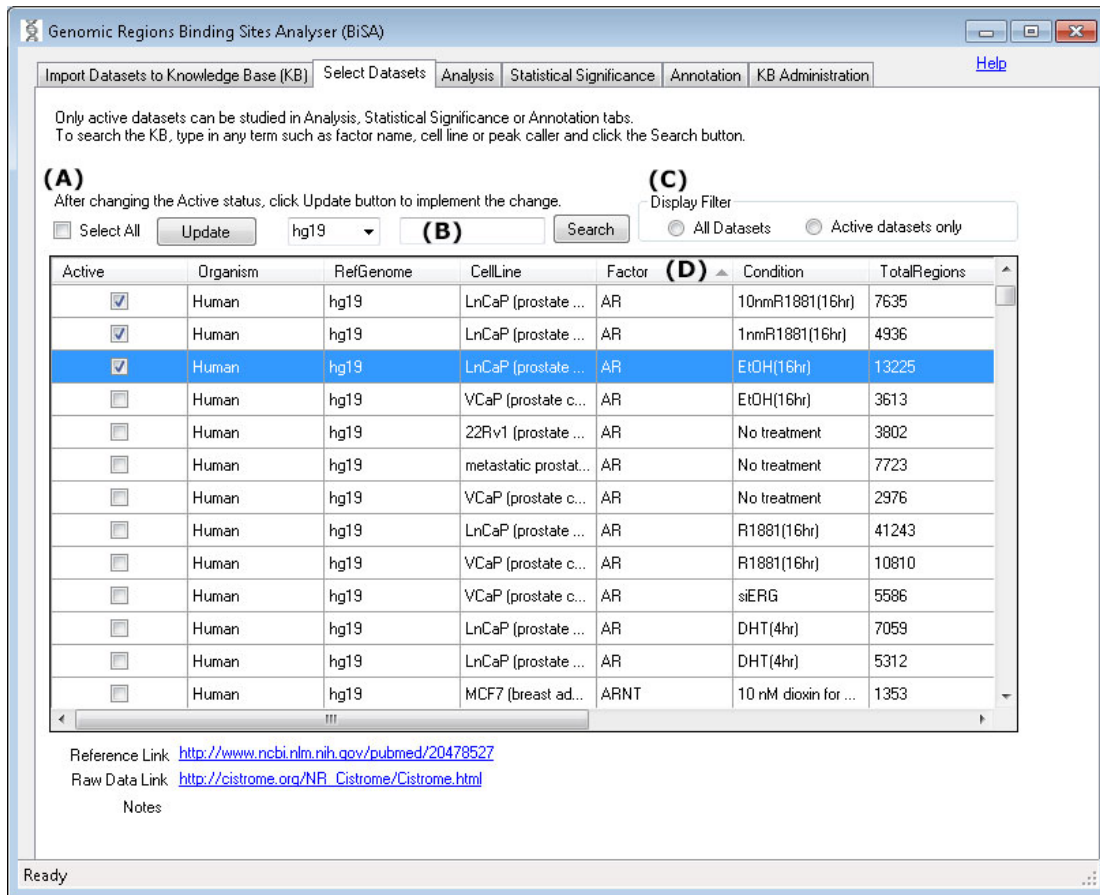


Figure 3-2: BiSA Select Datasets screen. This tab displays a list of all datasets in the KB, populated by default or as a consequence of uploading in Step-1. Clicking on the text of any row displays the reference link of the article, raw data link and notes, if any, below the table. Website addresses are hyperlinked to the websites/articles from where the data are obtained. (A) Changing the Active ticks and clicking on the Update button implements the selection. (B) Users can search the KB by organism, cell line, factor label, reference genome or peak caller.

The options a), b) and c) operate on all active datasets while options d), e) and f) are designed to work on two selected datasets. Ticking the “Extract both datasets, bp overlap and centre distance between the regions” for options d), e) and f) displays both Dataset-A and Dataset-B regions, bp overlap and distance between two sets. The number of base pairs (bp) either in common in two sets (set by a positive number) or separating two sets (set as a negative

number) can be specified, as can be the maximum allowed distance between the centres of two compared regions. Overlapping results can be visualized as Venn diagrams or saved to the KB or a tab-delimited text file (Figure 3.3, circled).

All analyses require setting minimum ‘bp overlaps’, however, specifying maximum distance allowable between two binding peaks or limiting results to a chromosome is optional. A positive value for minimum bp overlap would restrict results for regions that share the specified number of common base pairs. For instance, setting ‘bp overlap’ to 3 will report regions that have at least 3 bases in common. While studying TFs that compete for a specific DNA sequence or finding TFs that form a complex and bind to DNA, the minimum bp overlap can be set to 1 and maximum distance from the centre of two sets should be small, such as 50 bp. To study TFs that potentially bind close to each other a positive ‘bp overlap’ could be set keeping ‘maximum distance from the centre of two sets’ empty. To study TFs that do not overlap however tends to bind in proximity to each other, a negative value of minimum bp overlap can be assigned to report nearby regions. For example assigning a bp overlap of -100 will report nearby regions separated by up to 100 bases, in this case, a maximum centre distance should be specified. The analysis results section is a data grid that populates the results of the performed analysis (Figure 3.3, circled). Results can be saved in a tab-delimited text format, to allow further analysis in other software. Results can also be sorted by selecting any column heading. The Venn diagram button visualizes overlaps of a maximum of three activated datasets (Figure 3.3, circled). If there are more than three active datasets for drawing the Venn diagram then BiSA displays a warning (Figure 3.4A). Overlap statistics are displayed below the Venn diagram, which can also be saved to a file for later reference or figure preparation (Figure 3.4B). Overlapping or non-overlapping regions can be saved back to the KB (Figure 3.3) allowing them to go into downstream analysis and independent annotation.

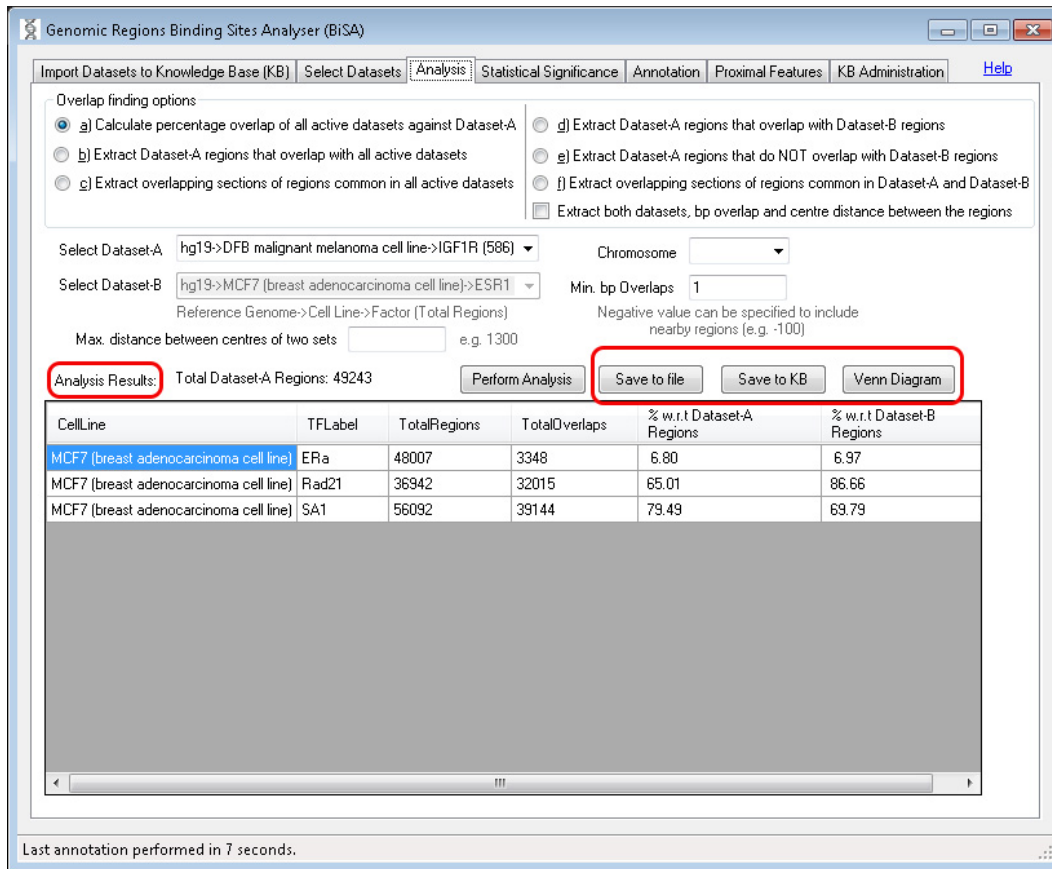


Figure 3-3: Analysis is the main overlap analysis tab of BiSA. BiSA offers six types of analysis: Overlap finding option a) reports overlap percentage with respect to the total Dataset-A regions and percentage with respect to the other active dataset regions. Overlapping or non-overlapping regions of Dataset-A can be extracted by options b), d) or e). Whereas, option c) or f) can be used to extract overlapping sections of regions common in all or two datasets. The results of overlap analysis type b), c), d), e) and f) can be saved back into the Knowledge Base by the 'Save to KB' button, allowing them to go into downstream analysis and independent annotation. Ticking the "Extract both datasets, bp overlap and centre distance between the regions" for options d), e) and f) displays both Dataset-A and Dataset-B regions, bp overlap and distance between the two sets.

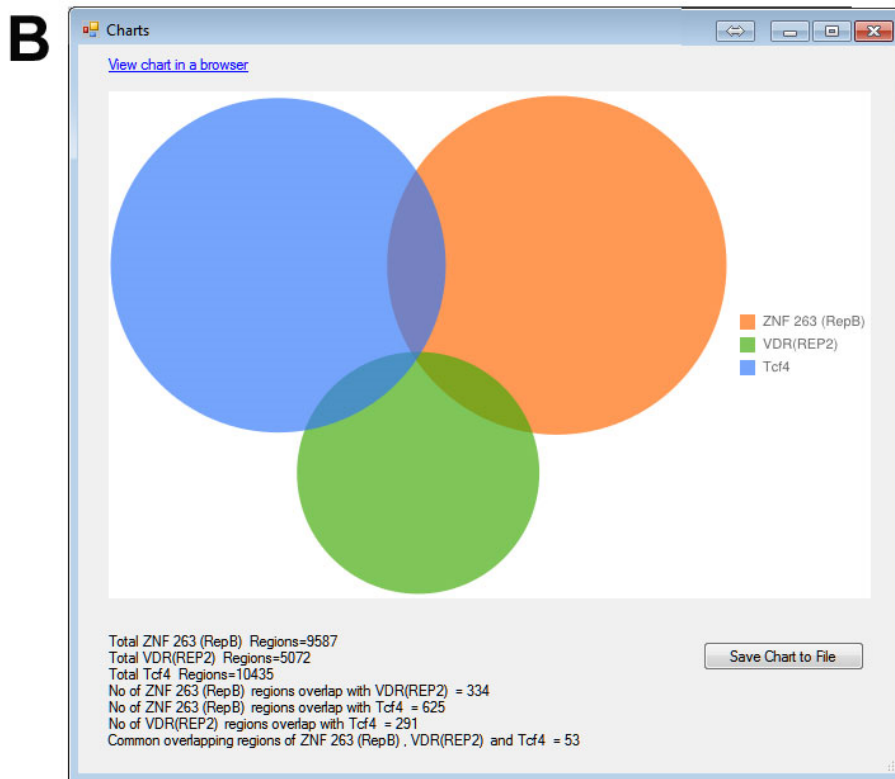
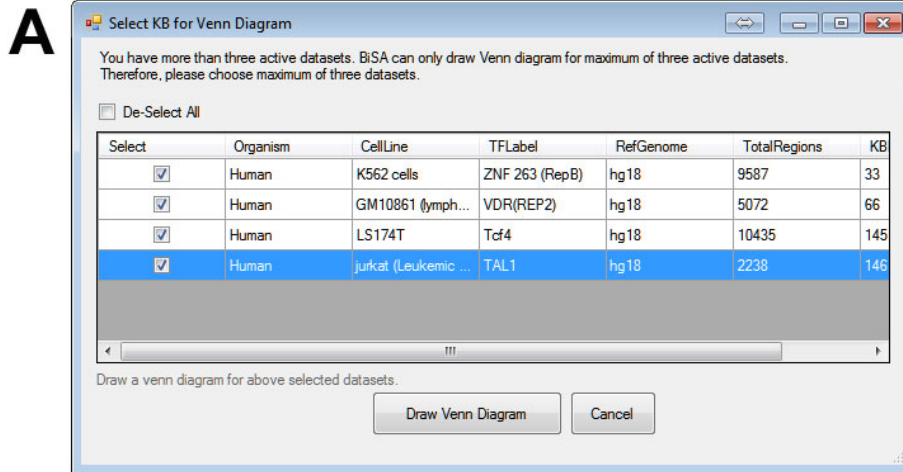


Figure 3-4: Venn diagrams in BiSA. BiSA can cross-compare a maximum of three active KB as a Venn diagram. (A) If there are more than three active datasets then a pop-up window appears that allows the investigator to select three datasets to be analysed. (B) Google Charts is used to draw Venn diagrams. The diagram can be saved as a high quality PNG file.

3.2.4. Statistical Significance

The number and location of TF binding regions discovered in a ChIP-seq experiment is influenced by experimental design, model used, sequencing depth and analysis approach. Therefore, this information is made available in as much detail as possible in BiSA, so that users can make judgements about the appropriateness of specific dataset comparisons. To determine the level of statistical significance of the degree of overlap in two datasets, the IntervalStats command line algorithm (Chikina and Troyanskaya, 2012) is implemented in a user friendly graphical interface. Active datasets to be compared are selected via two dropdown lists (Figure 3.5). Users can select one dataset as a query and the other one as a reference. IntervalStats only takes into account the regions that are within a defined domain dataset, representing the total available genomic area for binding. The results are saved as a tab-delimited text file with the regions from Dataset-A (query) and Dataset-B (reference), Dataset-A region size, the distance between them and the corresponding numerator and denominator used to calculate the p-value, which is saved as the last column. Once the IntervalStats tool finishes the process and the user closes the terminal window, BiSA calculates and displays an Overlap Correlation Value as described in the Section 2.13, which reflects the overall significance of overlap of the two datasets.

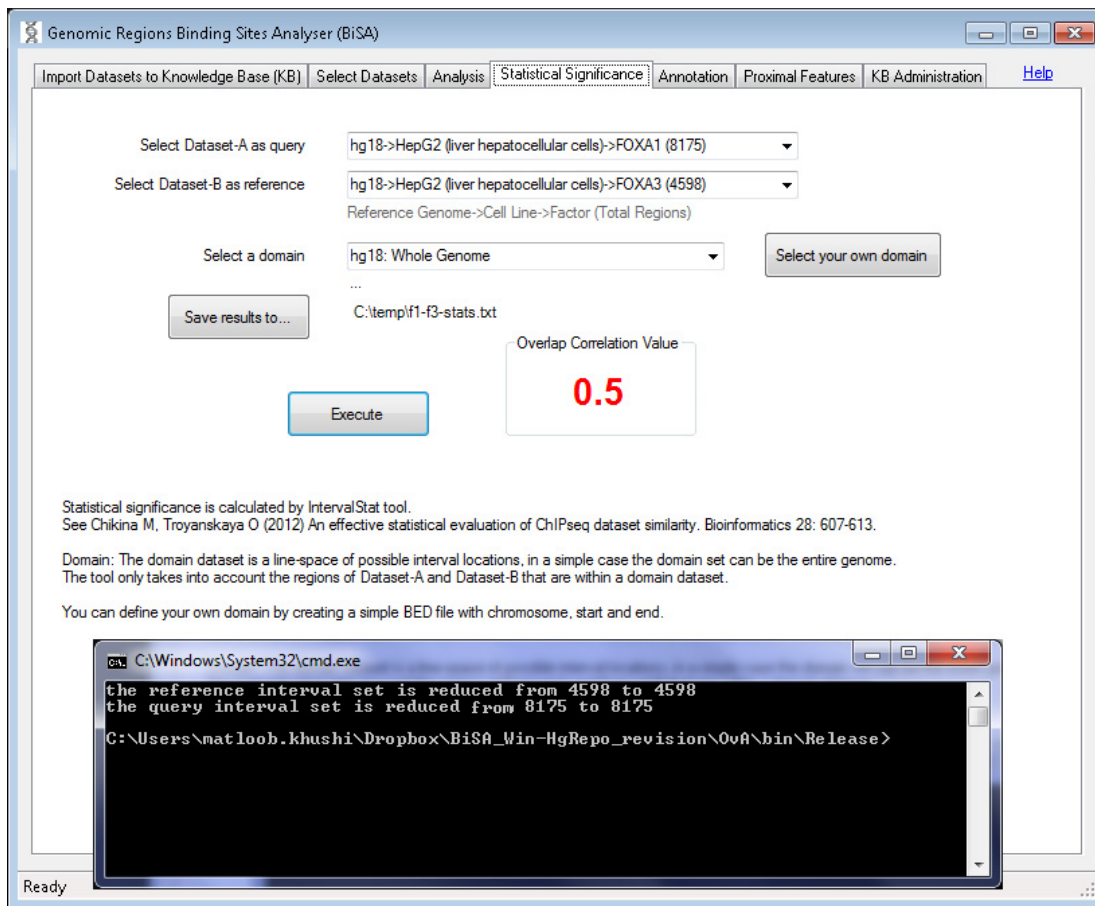


Figure 3-5: Statistical significance of overlapping regions. The statistical significance tab allows the user to determine the statistical significance of the extent of overlap of two sets of regions. Active datasets are loaded into two dropdown lists and the user selects one dataset as a query and the other one as a reference. Only regions of both datasets that are within the selected domain dataset are included in the calculation. Clicking the Execute button calls up a command-line window and executes the IntervalStat tool. The command-line window stays open to display the messages from the tool. When the terminal window is closed BiSA calculates Overlap Correlation Value of the two datasets.

3.2.5. Annotation

The annotation tab (Figure 3.6) allows the user to add nearby gene information to a selected set of binding regions. Users define maximum distances between binding peak and transcription start and end sites of nearby genes. The nearest gene per region or all genes

within the designated number of bp limits will be reported. Selecting “Load new genes” (Figure 3.6) allows custom gene definitions for additional organisms to be uploaded (Figure 3.7). The delete genes button allows the user to delete the custom uploaded definitions. We also calculate the distance between centre of region to TSS represented as bpTSS, whereas, distance from the transcription end site (TES) shown as bpTES. Therefore, a negative value in the bpTSS or bpTES column indicates that the region is upstream of the annotated TSS or TES respectively.

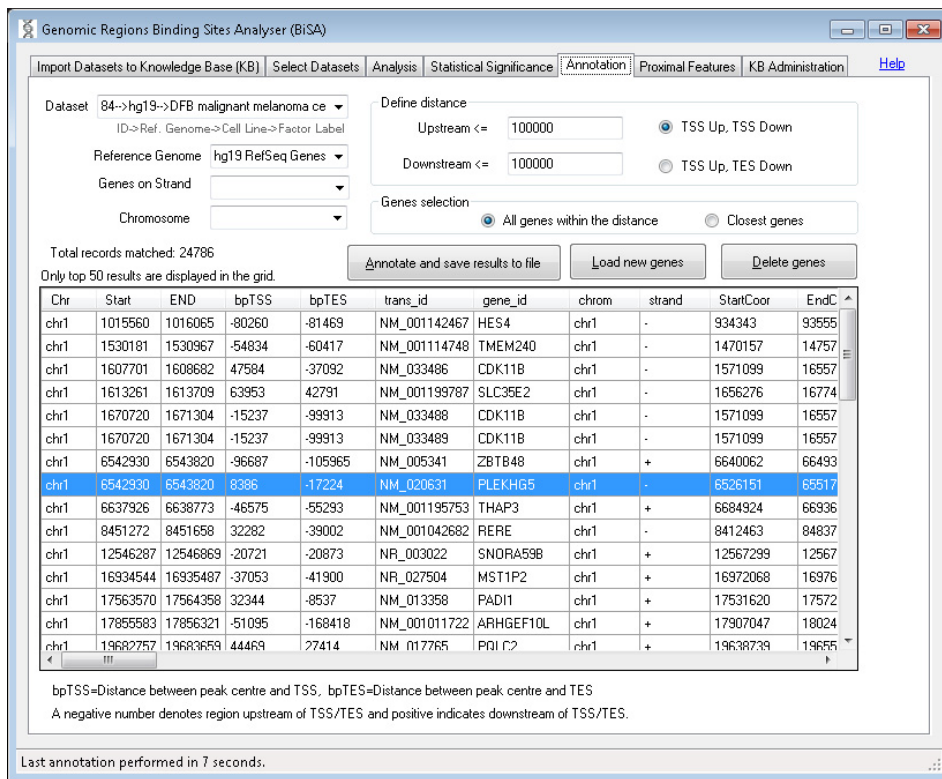


Figure 3-6: Gene annotation. The annotation tab allows the user to add gene information, from human and mouse reference genome assemblies, taken from the UCSC Genome Browser, to their data. This data can be saved in tab-delimited text format for further analysis in other software. Annotation can be limited to a chromosome and strand. Start and End co-ordinate columns for transcript (tx) and cDNA (cds) represent the numerically lower and higher value chromosomal coordinates for genes on both strands. A negative value in the bpTSS or bpTES column indicates that the region is upstream of the annotated TSS or TES

respectively. Therefore a region within a gene on the positive strand will have a negative bpTES value and a positive bpTSS value as for the region highlighted. Only the top 50 results are displayed in the grid, however, the full annotated dataset is saved in a tab-delimited text file which can be opened in Excel or other spreadsheet management software for further analysis. The delete genes button allows the user to delete custom uploaded definitions.

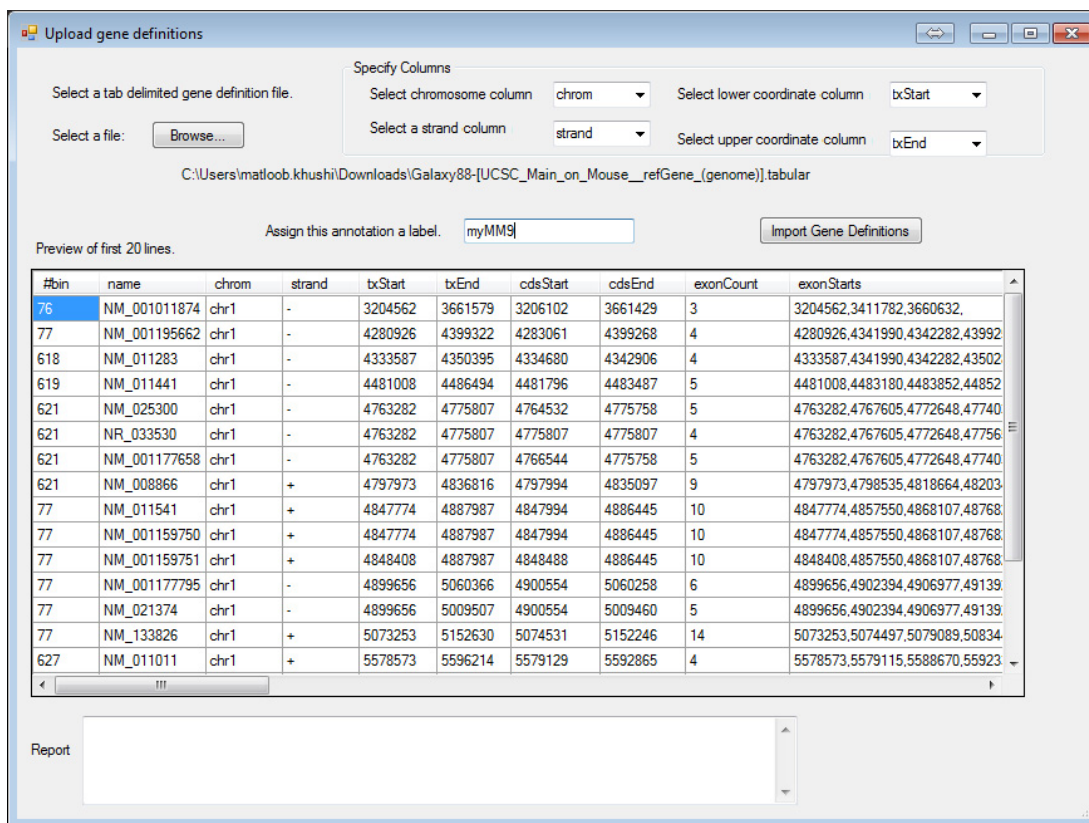


Figure 3-7: New gene definitions. New gene definitions may be uploaded in the software. The user must specify columns for chromosome, strand, lower and upper coordinates.

3.2.6. Proximal Features

This tab lets the investigator discover features that are in proximity to a gene of interest. The nearby genomic features can be discovered by specifying a locus, chromosome and position (Figure 3.8A) or a gene (Figure 3.8B). The gene can be searched by specifying an assembly such as hg19 and typing either the exact gene symbol or typing the first few letters of the gene name and pressing the Search button which brings up a list of matching genes. Once a

gene is selected, its chromosome, strand, TSS and CDS (Coding DNA Sequence) are displayed and the user can select whether the distance should be calculated from the gene TSS or CDS (Figure 3.8C). The base-pair distance between genomic features and the regions is calculated from the centre of the regions and can be set (Figure 3.8D). Selecting ‘all active datasets’ reports cell line, feature/factor and total regions found within the specified distance. If the user selects a single KB dataset then full details of all regions within the specified distance are reported which can then be saved back into the KB. All results can also be saved to a file.

3.2.7. Administration

From the Administration tab (Figure 3.9) users can delete a dataset, save selected data in a tab-delimited format, and view regions or region sizes. The distribution of region sizes over the dataset can also be listed or can be visualised as a histogram (Figure 3.9A). The Clean Up Database button (Figure 3.9B-circled) truncates transaction logs, to avoid an impact on software performance.

This screen lets you discover genomic features that in proximity of a gene of interest.

(B) Search a gene

Gene
 Assembly: hg19
 Gene symbol: BRCA1 [Search]
 Chromosome: chr17 Strand: -
 Distance from:
 Transcription Start Site (TSS) 41196311
 Coding Sequence (CDS) 41197694

(A) Or specify a locus

Locus
 Chromosome: [] Position: []

Feature discovery options
 Report factors/regions within: 100000 **(D)**
 Distance is calculated from the centre of a region.
 All active datasets: Factor and region count will be reported.
 Or select a dataset: 84-->hg19-->DFB malignant melanoma cell line-->IGF1
 Regions within the specified distance will be reported.

Results:

CellLine	Factor	Total_Regions
HepG2 (liver hepatocellular cells)	RAD21	14
MCF7 (breast adenocarcinoma cell line)	Pol II	10
DND41 (T cell leukemia)	H3K4me1	14
Nhek (epidermal keratinocytes skin)	H3k27me3	3
Nhek (epidermal keratinocytes skin)	H3k4me3	8
NB4 (hematopoietic stem cell)	H3K9K14ac	15
MCF7 (breast adenocarcinoma cell line)	T24	1
HepG2 (liver hepatocellular cells)	HNF4A	6
HepG2 (liver hepatocellular cells)	FoxA2	2

Last annotation performed in 7 seconds.

Figure 3-8: Proximal features. This tab allows users to search for genomic features located in proximity to a specific gene or genomic locus. Searching multiple datasets returns the numbers of binding sites for each factor identified. Selecting a single factor returns detailed binding region information for interactions in proximity to the gene or locus of interest.

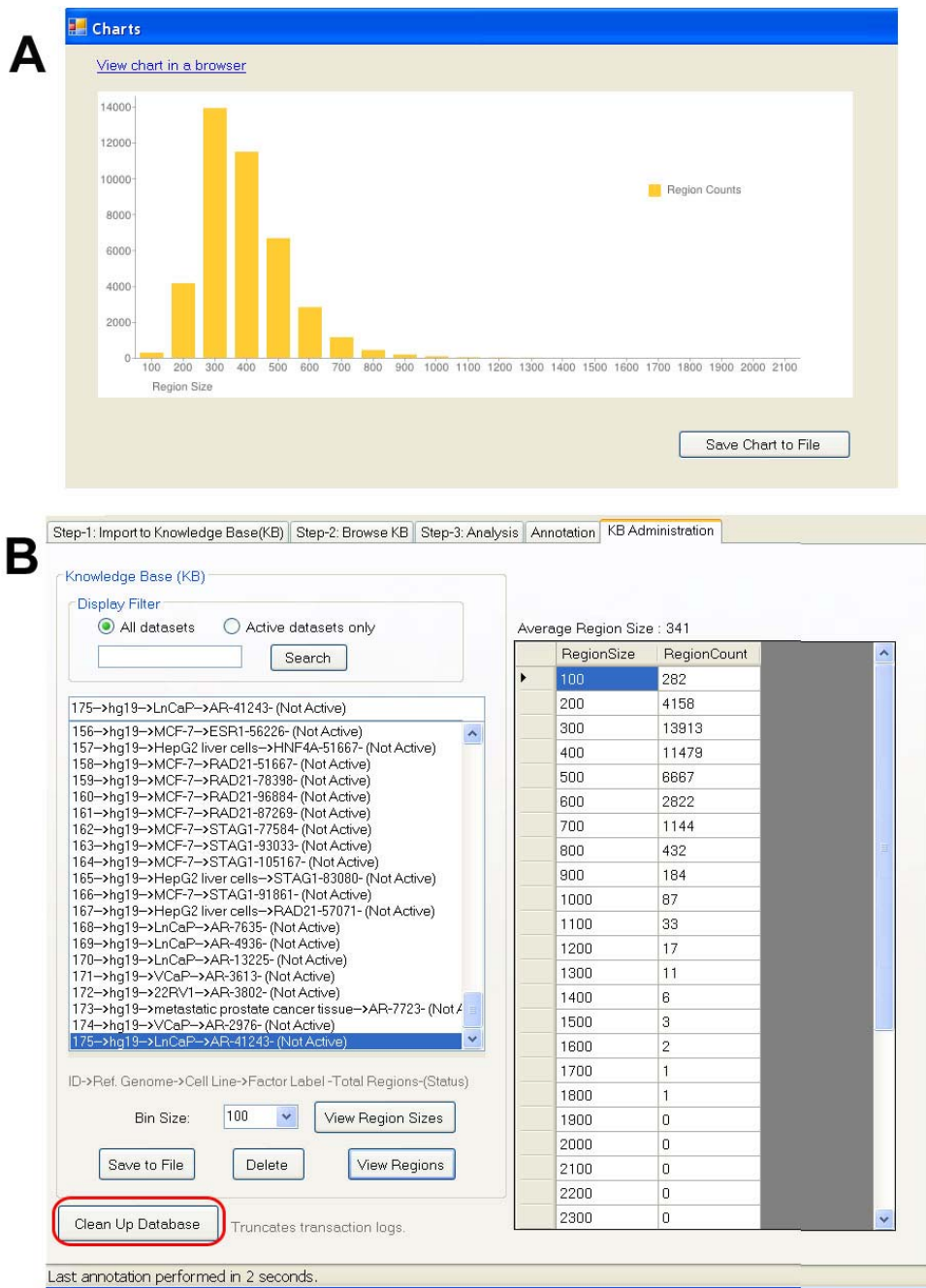


Figure 3-9: Administration of datasets. From the Administration tab users can A) view the distribution of region sizes over the dataset as a table and histogram, and B) delete a dataset, save the data in tab-delimited text format.

3.3. BiSA for Other Platforms

BiSA for Windows was designed to provide an easy desktop graphical user interface (GUI) for biologists to archive and analyse transcription factor binding sites and epigenetic

modifications. However, there are number of limitations in the desktop version identified as explained below:

1. Personal computers are not as powerful as computer servers due to their limited processing power and available memory. This limitation can be partly addressed by installing and configuring the database server (Microsoft SQL Server) on an enterprise server having larger expandable memory and processing. However, transferring datasets over the network can still be challenging and will be a bottleneck in an overall performance of the tool.
2. The second major limitation of the desktop version is difficulty in sharing of the results among other investigators. This means that an investigator has to manually save and move files across for combining analysis results performed by other researchers.
3. Thirdly many large bioinformatics facilities operate their computer server on Unix, Linux or Macintosh (Mac) type operating systems, therefore, adopting a Windows based solution institution-wide might not be a feasible solution for bioinformaticians.

Given this we also developed a web-based version of BiSA written for Unix-like operating environments. The web-based solution should be able to support other tools in the ChIP-seq and ChIP-chip analysis so that the output data from one pipe-line can be easily fed into the input of other tools and researchers should be able to share their results with other researchers anywhere in the world.

3.4. BiSA for Galaxy: Web Interface Overview

In the previous chapter I explained the basis on which it was decided to write BiSA for the Galaxy platform. Once successfully integrated, BiSA appears as a new tool-set on the left

Tools section (Figure 3.10) of the Galaxy main menu. Its order in the tools depends on the location of the BiSA XML code in `/galaxy-dist/tools_conf.xml` file. The Galaxy web interface is divided into three parts as shown by two red lines on Figure 3.10. The left vertical section shows all the available tools in Galaxy. Galaxy has a long list of pre-configured tools which can be searched using the search tools text box available on the top of the *Tools* panel (Figure 3.10A). The right vertical panel is called the *History* that shows all the uploaded datasets and results from a tool. Datasets are grouped into histories which can be named and the user can switch between different histories. The *Options* menu on the top of the *History* (Figure 3.10-C) provides the option of creating new history, deleting current history, switching between histories, retrieving or purging deleted datasets, sharing and publishing datasets. Middle section loads a web-form that collects all the values of parameters that are required to run a tool, for example the registration form (Figure 3.10).

Registering with Galaxy is mandatory for BiSA to maintain the privacy of each user and segregation of datasets from other users. This can be done by choosing the Register option in the User menu and completing four form fields: email, password, confirm password and public name (Figure 3.10). Password needs to be at least six characters long. A public name is not mandatory; this is the name that is shown to other users when the user shares datasets with other users or the public if they choose to. Once a user completes the registration form they are automatically logged in and the email address and password can be used for future logins.

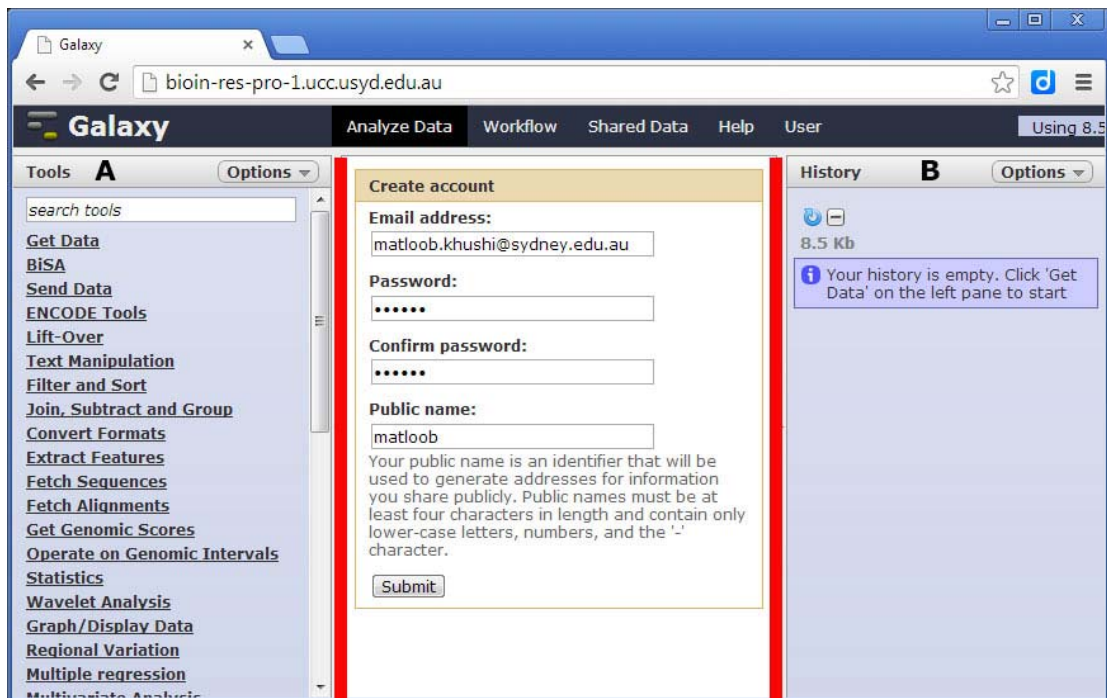


Figure 3-10: BiSA for Galaxy web interface overview. The web interface is divided into three vertical parts, A) the left section is called the Tools, B) the right section is called the History and the middle section loads the web-form for a tool such as the registration form.

3.4.1. Importing Datasets into BiSA

BiSA for Galaxy is a web-application where all processing is performed on the web server therefore a client can be anywhere in the world. Importing datasets into BiSA datasets is performed in two steps: datasets are first uploaded to the web-server and then imported into the BiSA database. The data can be uploaded to the web-server by any one of the following three methods:

1. Direct browser upload
2. FTP (File Transfer Protocol)
3. Specifying a URL (Uniform Resource Locator)

Direct browser upload is a good option for small data files of no more than a few Megabytes. The maximum theoretical limit of browser upload is 2GB, however, browser upload is slow for files that are even as large as few hundred Megabytes. A second disadvantage of using browser-based upload for large files is that the protocol of the communication is usually HTTP (Hyper Text Transfer Protocol) which is an unreliable protocol if compared to FTP (File Transfer Protocol) (Singh et al., 2013; Kurose and Ross, 2012). This means that if 90% of the file is sent and then there is a failure, upload has to be restarted from the beginning unless special flash-based, java-based or javascript-based unloaders are being used. Due to these issues, I have setup the server to accept SFTP (Secure File Transfer Protocol) uploads for larger files. Uploading files using a URL is another useful option when data is hosted on a remote server and a direct link to the data is known.

Datasets can also be imported directly from 23 recognised genomic facilities such as UCSC Table Browser and BioMart.

A file can be uploaded through the browser by selecting the Upload File option in Get Data menu (Figure 3.11-A). The file is selected by clicking on the Choose File button (Figure 3.11-B), the file format and the reference genome such as hg19 can also be specified. Once the file is uploaded, the file name becomes its label and is shown in the current history as a green highlighted text (Figure 3.11-C).

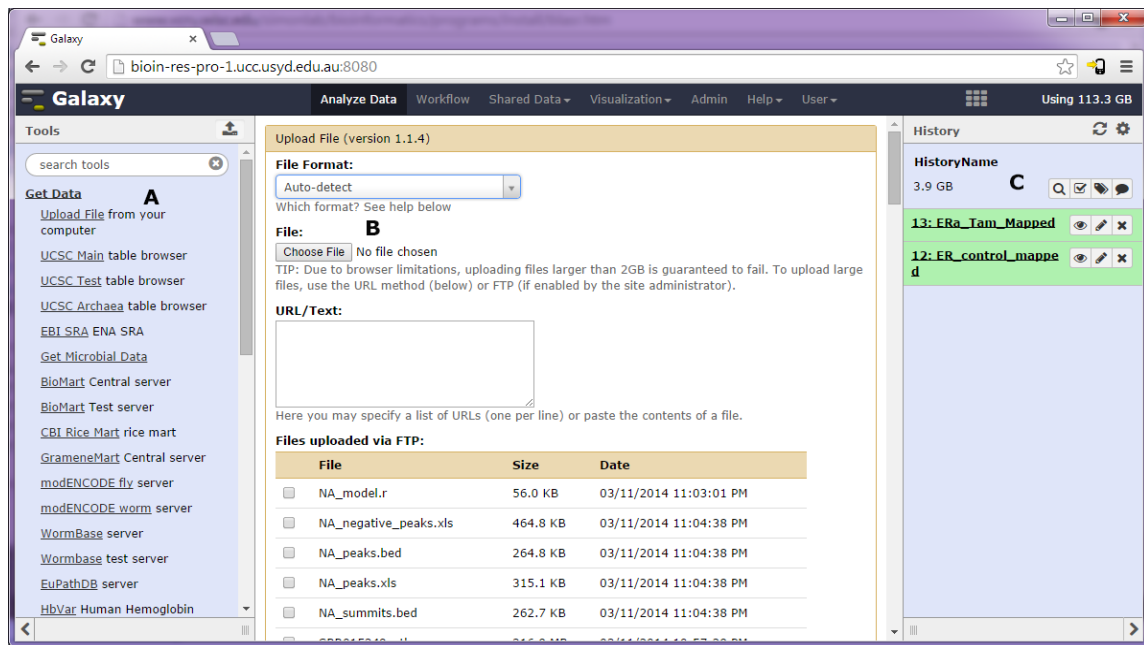


Figure 3-11: Uploading data into the Galaxy. A) Data file first needs to be uploaded into the web-server using Upload file option under Get Data menu item before it can be imported into the BiSA database. B) File to be uploaded can be chosen from local system. C) Uploaded file displays in history.

More information about the file can be seen by clicking on the file name (Figure 3.12). The expanded file information shows various options, a preview of data in the file, the total number of genomic regions in the file, and the file format. The various attributes of the file such as file format, genome, chromosome and coordinates columns can be changed by clicking on the pencil icon (Figure 3.12-circled).

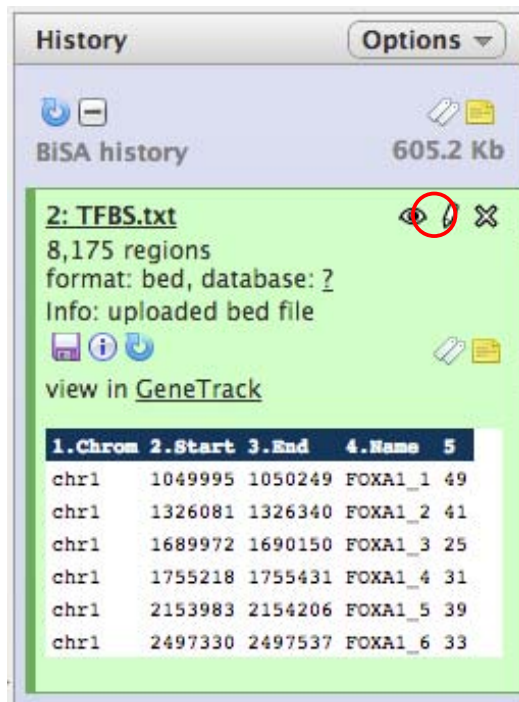


Figure 3-12: Clicking on the file name label reveals dataset information.

It is important that Galaxy recognises the uploaded genomic regions file as either BED or GFF file format. Galaxy usually can automatically detect the file format, however, it is a good idea to specify the format manually to avoid any misinterpretation by the Galaxy script.

Once the dataset is uploaded into Galaxy the dataset can be imported into BiSA by clicking on the 'Import Dataset to KB' option under the BiSA menu (Figure 3.13, circled). The file format must either be BED or GFF, but the file extension does not matter, for example in Figure 3.13 the file extension is TXT, however, its format is recognised as BED format therefore it automatically appears in the BiSA import dropdown box. The import datasets screen also provides text boxes to specify organism, reference genome, cell line and a unique data label.

BiSA also records the email address of the user who uploads the dataset to identify owner of the datasets. Users cannot analyse other users' datasets.

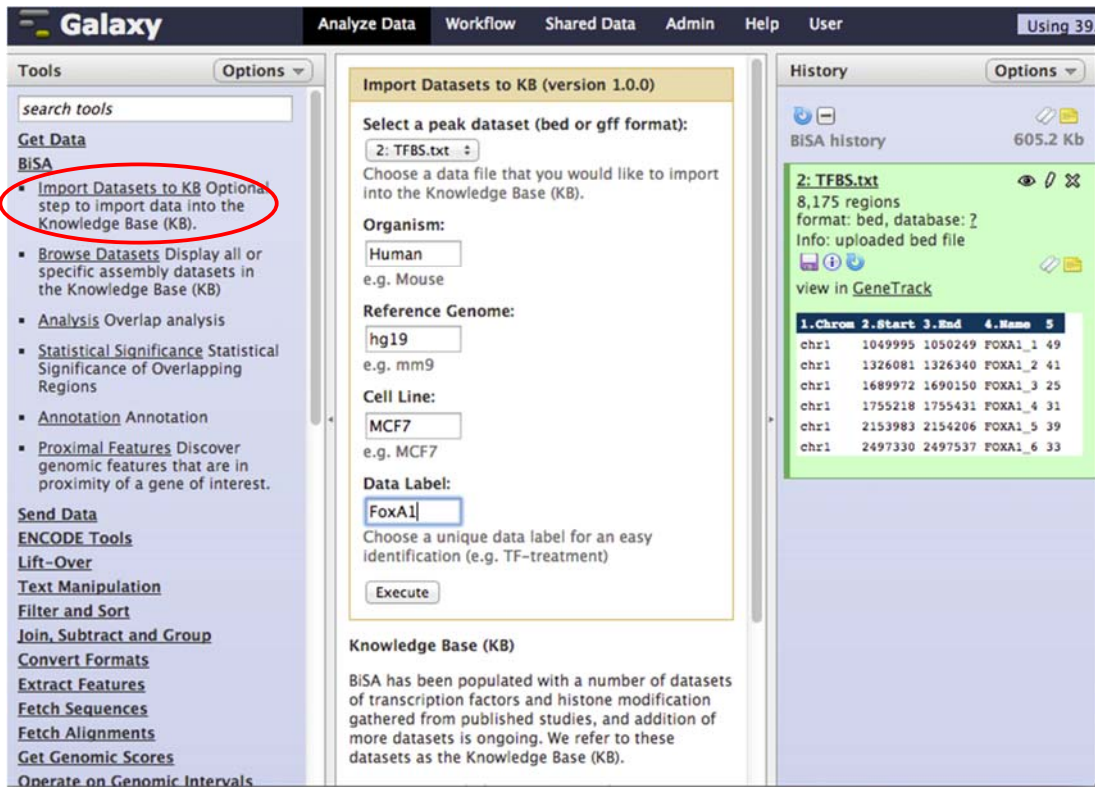


Figure 3-13: BiSA Import Datasets to KB (circled) automatically populates BED or GFF formatted datasets from the current history in the dropdown.

Clicking the Execute button adds the job to the queue waiting to be processed, and a new dataset is added in the History, shown grey in the Figure 3.14. Once the processing of the data starts the colour of the dataset is changed to yellow and when the job finishes its colour is changed to green. If the job fails for any reason the colour turns red. These colours help in identifying the current status of dataset processing. All details of the dataset and the email of the logged-in user are saved in the 'kbdetails' table while the genomic regions are saved in the 'kbsites' table. The email address identifies the owner of the database.

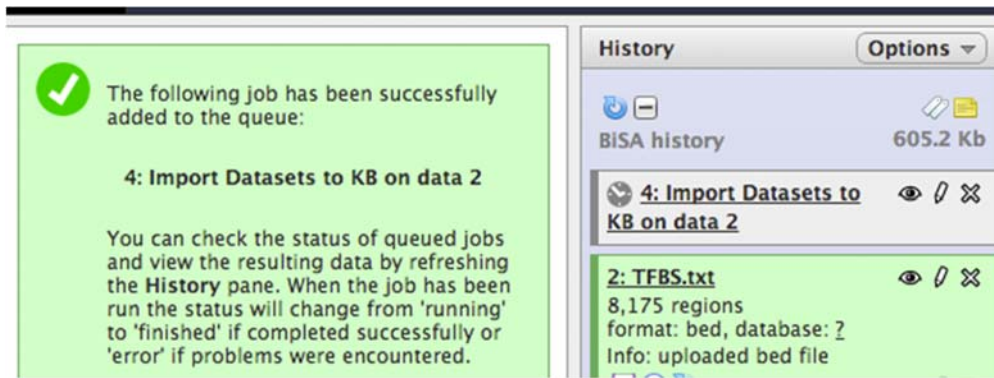


Figure 3-14: Submitting Import Datasets to KB form adds a new job in the queue shown in grey in the History. The colour of the dataset changes to yellow when processing and finally turns green when the processing successfully finishes. Failed jobs are shown in red.

3.4.2. Browse Datasets

Unlike BiSA for Windows there is no requirement to activate datasets in Galaxy. The *Browse Datasets* screen lets the user browse the available datasets to discover information about the datasets available in the knowledge base (KB). The dataset information fields are concatenated by a '→' sign starting with 'Default' or the seven initial letters of the email identifying the owner of the dataset. The 'Default' keyword represents that all users can use the dataset in any analysis, however, other datasets can only be used/analysed by their owners. Ideally BiSA should not display datasets belonging to other users, however, this is not possible at this stage because of a technical limitation in Galaxy. Hence, only labels of dataset are displayed and if users try to perform any analysis on a dataset that does not belong to them then BiSA displays an error message.

In *Browse Dataset* the datasets can also be viewed based on hg19, hg18, mm9, or mm8 or other genomes (Figure 3.15).

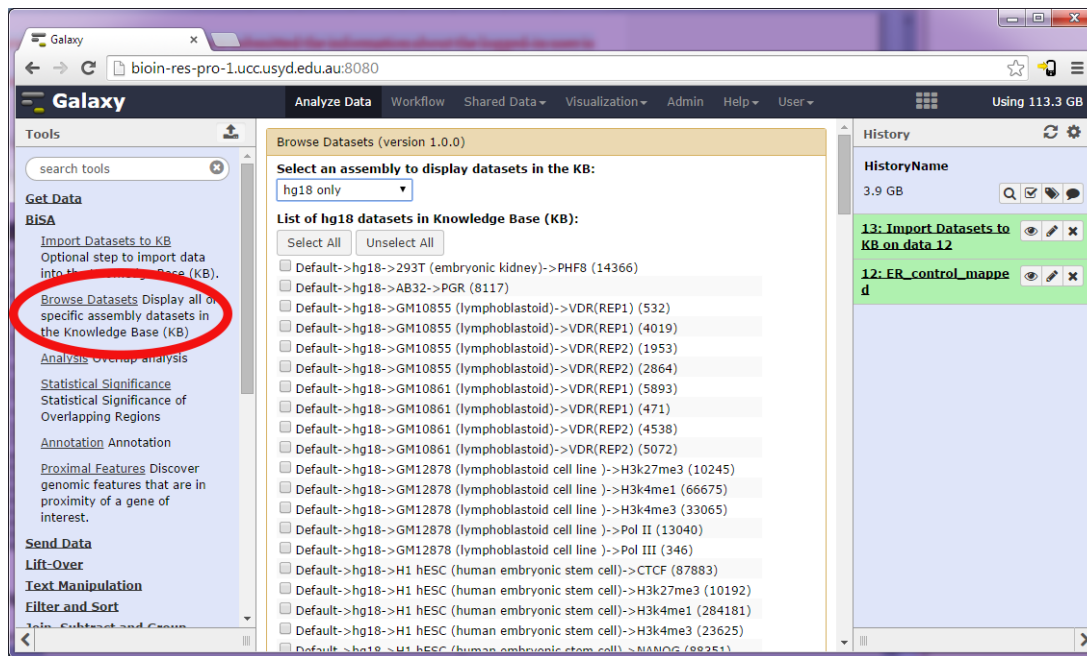


Figure 3-15: *Browse Datasets* can be used to browse the dataset information available in the Knowledge Base (KB).

3.4.3. Analysis

Analysis is the main screen where overlapping and non-overlapping datasets can be studied. Like BiSA for Windows, users can set a positive ‘minimum base pair overlap’ number for extracting regions that share common bases or a negative number to include regions that are nearby. Users can only analyse default datasets loaded as KB or datasets belonging to them, otherwise analysis fails with a message on the *History*. Similarly analysis can be restricted to a chromosome or maximum distance between centres of the regions. Unlike the Windows version, it is not required that datasets are activated for analysing, rather all datasets are populated in two dropdown lists filtered based on the selected assembly (eg hg18) (Figure 3.16).

Analysis (version 1.0.0)

Minimum bp overlap:

 The numbers of base pair (bp) in common in two sets is set by a positive number or away from either end can be set by a negative number.

Maximum distance between centres of two sets (optional):

Restrict analysis for a chromosome (optional):

 e.g. chr1

A Save results:
 ▾
 Save to KB will only works for Output Option where result returns in the form of genomic regions`.

B Select an assembly to display datasets in the KB:
 ▾

Select Dataset-A:
 ▾
 Default KB/Owner User->Reference genome->Cell line->Label (Total Regions)

Select Dataset-B:
 ▾
 Default KB/Owner User->Reference genome->Cell line->Label (Total Regions)

Dataset Series

C

Output option:

- a) Extract Dataset-A regions that overlap with Dataset-B and Dataset Series (if any).
- b) Extract Dataset-A regions that do NOT overlap with Dataset-B regions (ignores Dataset Series).
- c) Extract Dataset-A overlapping pieces of regions common in Dataset-B and Dataset Series (if any).
- d) Calculate percentage overlap of Dataset-A with Dataset-B and Dataset Series.
- e) Draw a Venn Diagram of overlaps for maximum of three datasets.
- f) Calculate average region sizes for Dataset-A (round up to two digits).
- g) Delete Dataset-A (you can only delete your own uploaded datasets).

Figure 3-16: Analysis is the main screen where overlapping and non-overlapping datasets can be studied. A) Option to save results in the Galaxy history or back into KB. B) Option to filter datasets based on an assembly. C) Option to add more datasets for analysis.

The result dataset is saved on the *History* panel as described previously. However, if the result is required to be saved back into the Knowledge Base (KB) then the Save results drop down value should be changed to ‘Save result back to KB’ (Figure 3.16-A), selecting this

option reveals a new text box to enter the label for the dataset (Figure 3.17). Saving results back to KB only works for analysis options a), b) and c) where result datasets are genomic regions. More than two datasets can be analysed by clicking on ‘Add new Dataset Series’ (Figure 3.17-C), which adds a new dropdown menu populated with all datasets of the selected assembly (Figure 3.17).

Save results:
 ▾
 Save to KB will only works for Output Option where result returns in the form of genomic regions.

Select an assembly to display datasets in the KB:
 ▾

Select Dataset-A:
 ▾
 Default KB/Owner User->Reference genome->Cell line->Label (Total Regions)

Select Dataset-B:
 ▾
 Default KB/Owner User->Reference genome->Cell line->Label (Total Regions)

Dataset Series

Dataset Series 1

Select dataset:
 ▾
 Default KB/Owner User->Reference genome->Cell line->Label (Total Regions)

Figure 3-17: A section of Galaxy analysis screen. Result can be saved back to knowledge base database by changing ‘Save results’ dropdown value to ‘Save result back to KB’. More than two datasets can be studied by adding a new Dataset Series.

Analysis options are largely similar to BiSA for Windows. However due to Galaxy’s web-based technical environment features are designed differently. For example in the Windows version results are shown in a grid and are not saved on disk or in KB until relevant button

options are selected. In Galaxy all results are saved in the Galaxy *History* by default and are not shown on screen until explicitly viewed. Here I briefly describe the analysis options and their difference with the Windows version if any.

a) Extract Dataset-A regions that overlap with Dataset-B and Dataset Series (if any):

This option extracts the overlapping regions of the selected Dataset-A against Dataset-B and if more datasets are also selected through adding Dataset Series then the overlapping regions of Dataset-A and Dataset-B are checked for overlapping regions in the Dataset Series. The results can be saved back to KB by changing the ‘Save result’ option (Figure 3.16-A).

b) Extract Dataset-A regions that do NOT overlap with Dataset-B regions (ignores Dataset Series).

This option extracts the non-overlapping regions of Dataset-A. This option is designed to work for only two datasets in selection, therefore, disregards any dataset in Dataset Series if added.

c) Extract Dataset-A overlapping pieces of regions common in Dataset-B and Dataset Series (if any).

This option extracts the Dataset-A overlapping sections of regions common in all selected datasets.

d) Calculate percentage overlap of Dataset-A with Dataset-B and Dataset Series.

Like the Windows version, this option calculates pair-wise total number of overlaps and its respective percentage. The result is generated as an HTML file saved on the *History*. To view the results the ‘eye’ icon on the History dataset is clicked (Figure 3.18-circled) and the result is shown in the main window.

Percentage Overlap of KB

KB-A Label = FOXA1
Having Total Regions = 8175

Here is the Result of KB-A Overlapping with the selected KBs.

Cell line	TF Label	Total Regions (KB)	Total Overlaps	%age w.r.t. KB-A regions	%age w.r.t. this KB regions
Human liver	FOXA2	8023	2518	30.80	31.38
HepG2 (liver hepatocellular cells)	FOXA3	4598	2939	35.95	63.91
HepG2 (liver hepatocellular cells)	CEBPA	30413	3215	39.32	10.57

w.r.t. = with respect to

Figure 3-18: HTML output of Galaxy analysis option. Option d) extracts Dataset-A overlapping sections of regions common in the all datasets and saves as HTML file in the History which can be viewed by clicking on the eye icon (circled).

e) Draw a Venn diagram of overlaps for a maximum of three datasets.

This option calculates the total number of overlapping regions among the three datasets and shows the overlaps graphically by drawing a Venn diagram. The Venn diagram is saved as a HTML file on the *History* which can be seen by clicking on the eye as described above. The Venn diagram in the Figure 3.19 shows the overlapping of three hg18 datasets FOXA1, FOXA2 and FOXA3 available in the KB. The diagram image can be saved locally by right clicking on the image choosing ‘save image as’.

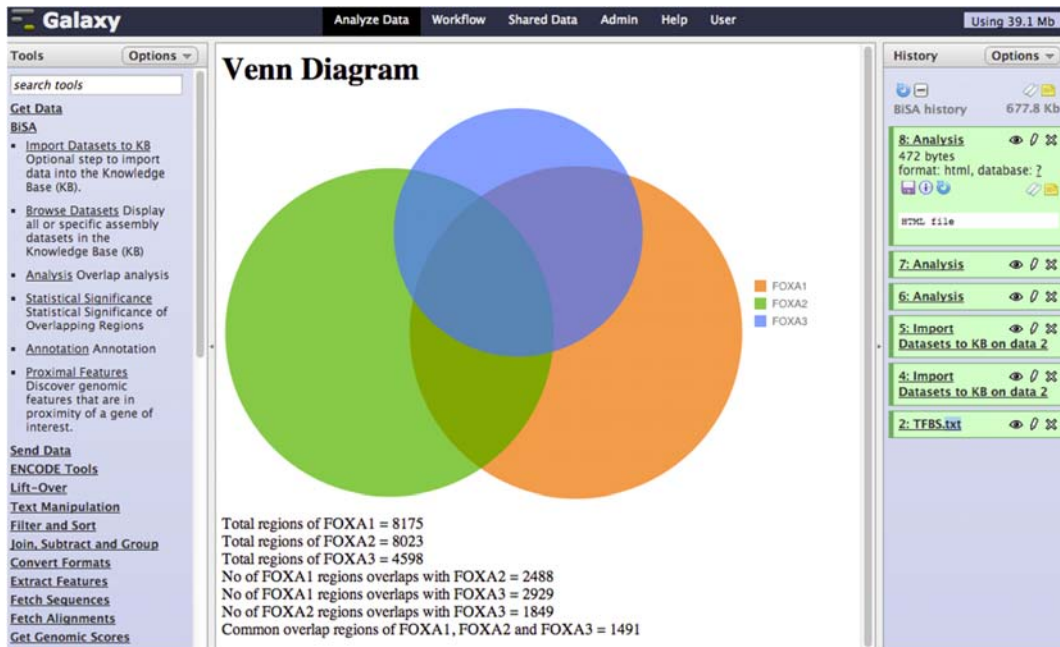


Figure 3-19: Venn diagram is drawn as an HTML file and overlapping numbers used in the drawing are also shown. The Venn diagram shows the overlapping of three hg18 datasets FOXA1, FOXA2 and FOXA3 available in the KB.

f) Calculate average region sizes for Dataset-A (round up to two digits).

This option again generates an HTML file for the bin size of 100 and total number of regions counted in a descending order. Figure 3.20 shows the region sizes calculated for the Motallebipour et al. (Motallebipour et al., 2009a) FOXA1 dataset. This reveals that ~90% of FOXA1 (7210 out of 8175 total regions) binding sites range from 100 to 400 bases.

g) Delete Dataset-A (you can only delete your own uploaded datasets).

In the Windows version the option to delete a dataset is provided under Administration tab, however, in the Galaxy version I have not developed a separate Administration link, therefore, a deletion option is provided here. When this option is executed BiSA checks the ownership of the dataset and if it matches with the logged-in user then the delete SQL

command is executed. Users cannot delete the default populated BiSA Knowledge Base (KB) or datasets that do not belong to them.

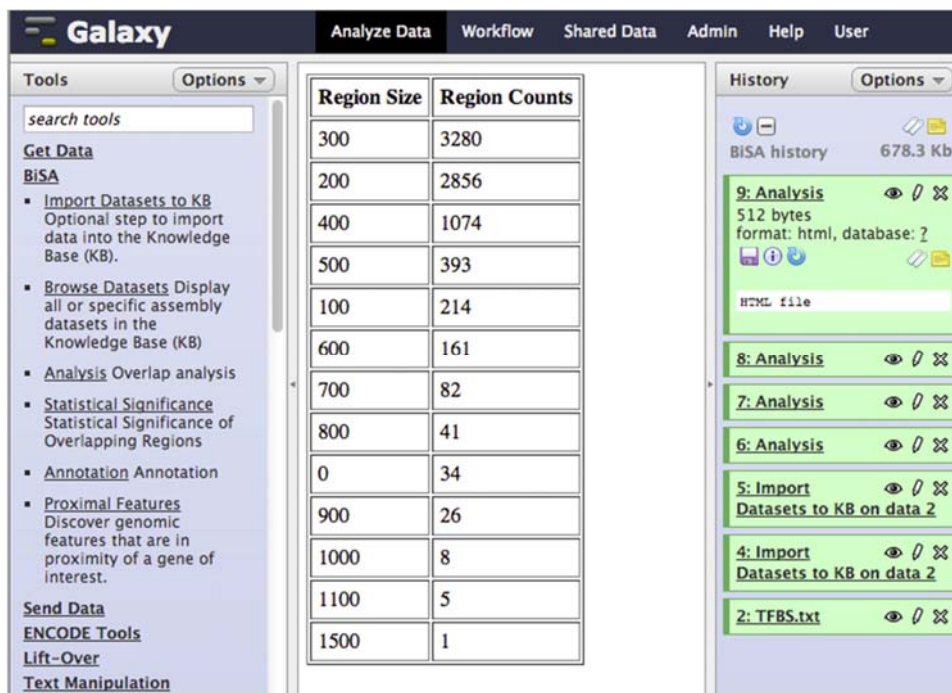


Figure 3-20: Region sizes are calculated for a bin size of 100 and presented as a HTML file saved on the History.

3.4.4. Statistical Significance

Statistical significance is determined in exactly the same way as in BiSA for Windows except that the Overlap Correlation Value and the output file are displayed on the *History* (Figure 3.21). As explained in Chapter 2, the p-values are calculated by using the IntervalStat tool (Chikina and Troyanskaya, 2012) InvervalStats tool calculates the p-value of each query against the closest reference region, this information for all the regions can be downloaded by clicking the save icon in the *History* (Figure 3.21-circled).

The screenshot displays the Galaxy web interface. On the left, the 'Statistical Significance (version 1.0.0)' tool is shown with the following configuration:

- Assembly: hg18
- Dataset-A: FOX A1 (8175)
- Dataset-B: FOX A3 (4598)
- Domain type: Select a built-in domain
- BiSA built-in domain: hg18: Whole Genome

 The 'Execute' button is visible at the bottom of the tool panel. On the right, the 'History' panel shows a list of jobs. Job 186, 'Statistical Significance', is highlighted in green. It shows 8,175 lines of output in tabular format with an overlap correlation value of 0.50. Below this, a table of genomic coordinates is displayed:

1	2	3
chr1: 1049995: 1050249	chr1: 1050036: 1050235	25
chr1: 1326081: 1326340	chr1: 1326099: 1326343	26
chr1: 1689972: 1690150	chr1: 1689918: 1690134	17
chr1: 1755218: 1755431	chr1: 1704307: 1704562	21
chr1: 2153983: 2154206	chr1: 2153947: 2154169	22
chr1: 2497330: 2497537	chr1: 2497307: 2497517	26

 The history panel also shows a circled save icon (floppy disk) next to job 186, indicating that the output file can be saved. Other jobs like 'Statistical Significance' (185) and 'Annotation' (172) are also visible in the history list.

Figure 3-21: The statistical summary, the overlap correlation value and the output file is displayed on the History. The output file containing the p-values of all regions can be saved by clicking on the save icon in the History (circled).

3.4.5. Proximal Features

This feature is the replication of the Windows version where transcription factor binding sites or histone modifications that are near to a gene or locus of interest can be discovered by specifying a distance within which the features are to be extracted. In Galaxy version the web-interface is different from the Windows version, by default the form (Figure 3.22-A) loads all hg19 genes, however, other genomic annotation datasets can be loaded by changing assembly to other genomes. The distance is calculated from the transcription start sites (TSS) of the genes. Changing 'Find genomic features from' gene to locus hides all genes and displays two text boxes for chromosome and position (Figure 3.22-B). By default the regions are counted for each dataset of a selected assembly (such as hg19) and the output shows the

cell line, factor label and the counted regions within the distance. However if actual regions are required to be extracted then the option should be changed to ‘Select dataset to extract nearby regions’ and a dataset is selected (Figure 3.22-B).

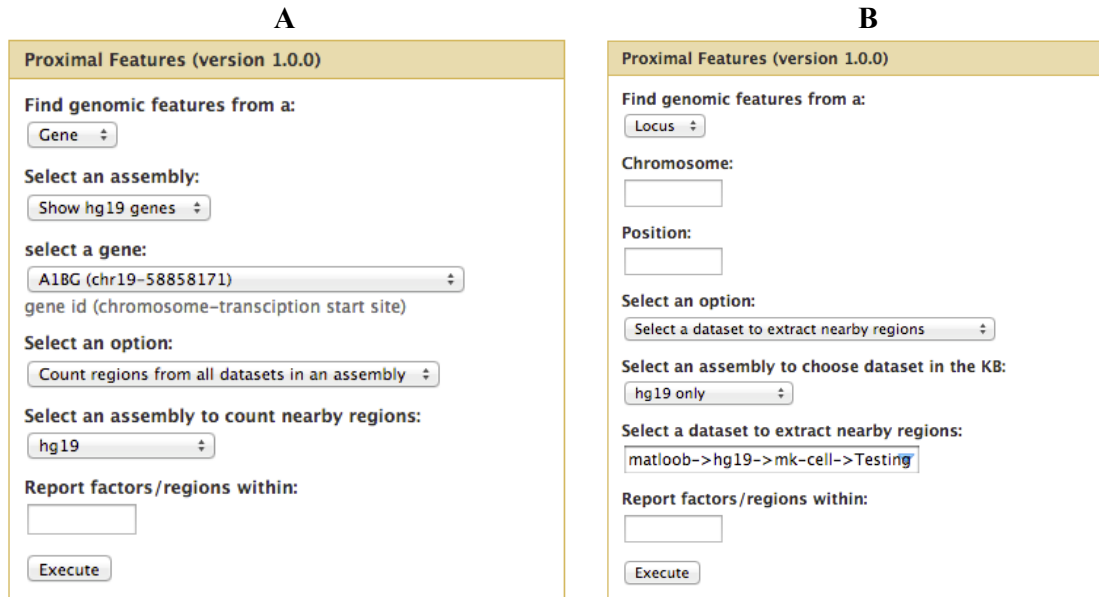


Figure 3-22: Genomic features can be extracted with a given distance. A) All genes for a selected assembly are loaded and distances are calculated from the transcription start sites (TSS). B) Selecting ‘Finding genomic features from a locus’, genes are hidden and two text boxes for chromosome and position are displayed. Actual regions can be extracted by selecting a dataset.

3.4.6. Annotation

The datasets saved in the Knowledge Base (KB) can be annotated for the nearest genes and distance from transcription start sites (TSS). Similar to the BiSA for Windows upstream and downstream distances are specified and distances can be calculated from only the TSS or downstream distance can be calculated from the transcription end sites (TES). All genes with the distance can be reported or can be restricted to closest genes, chromosome or strand.

Output file is tab-delimited text file linked on the *History*.

3.5. BiSA Application Example

To test the utility of BiSA, we studied six hg18 datasets available in the KB, transcription factors FOXA1 versus FOXA3 (Motallebipour et al., 2009a), CTCF versus SA1 (Schmidt et al., 2010) and ZNF263 (Frietze et al., 2010) versus c-Fos (Raha et al., 2010). The forkhead family of pioneer factors (FOXA1, FOXA2 and FOXA3) play important roles in early development to metabolism and homeostasis in adults, and are required for regulation of liver specific genes (Lee et al., 2005; Motallebipour et al., 2009a; Friedman and Kaestner, 2006). Their DNA-binding domains are highly conserved from yeast to mammals, and there is evidence for cooperative function between the family members (Motallebipour et al., 2009a; Soccio et al., 2011; Friedman and Kaestner, 2006). FOXA factors are pioneer factors due to their ability to bind condensed chromatin and reposition nucleosomes, allowing the binding of other factors (Friedman and Kaestner, 2006). These TFs work together in complex ways to regulate transcription, and co-location of binding sites of these factors have been extensively studied in the HepG2 cell line (Motallebipour et al., 2009a; Wallerman et al., 2009). Here we demonstrate the application of BiSA by investigating the overlap of binding sites for FOXA1 (8175 regions) and FOXA3 (4598 regions) (Motallebipour et al., 2009a) in HepG2 cells. We have also examined the MCF7 cell line datasets of Schmidt et al. for the overlap between CTCF and the cohesin complex component SA1 which are known to collocate on DNA (Schmidt et al., 2010). In addition we also studied two non-related transcription factors c-Fos (18211 regions) (Raha et al., 2010) and ZNF263 (4426 regions) (Frietze et al., 2010) in the K562 (erythromyeloblastoid leukemia) cell line.

BiSA overlap analysis of FOXA1 and FOXA3 with at least 1 bp in common reveals that 2929 FOXA1 regions overlap with FOXA3. On the other hand, 2937 FOXA3 regions overlapped with FOXA1, BiSA reported 5,246 unique FOXA1 binding sites and 4598 unique FOXA3 binding sites. To investigate further, when we extracted the overlapping common sections of

the regions of two transcription factors the overlapping number increased to 2,939 regions which shows that some regions of the two datasets overlap more than one region of the other dataset. To show this interaction graphically we drew a Venn diagram (Figure 3.25-A). The Venn diagram shows 2,939 common sections between the two datasets due to which the sum of common sections and unique regions do not equal to total regions of the dataset. We saved the overlapping sections back into the KB.

‘View Region Sizes’ under the Administration tab is used to draw a histogram of region sizes using bin size 100 (Figure 3.25-B). The histogram, showing the distribution of overlapping region sizes, reveals that ~99% of overlaps exceed 200 bases and there are more than 1600 regions that have at least 300 bp in common between the two datasets. Similarly the overlap analysis (39,568 common regions) of CTCF (49,243 regions) and SA1 (56,092 regions) is drawn as a Venn diagram and overlapping sections are represented in a histogram (Figure 3.25-C,D). Similar to the FOXA1-FOXA3 example, the number of common overlapping sections (39586) is greater than the total number of unique overlapping CTCF binding sites (39,144) due to the fact that a subset of regions overlap multiple regions in the comparison dataset. By contrast, when the unrelated transcription factors, c-Fos and ZNF263, are compared, just 559 overlaps are detected as expected for unrelated TFs. A Venn diagram showing the dataset overlap and a histogram summarizing the overlaps are drawn (Figure 3.25-E,F). We also observed that in three comparisons >94% of the overlapping sections are >200 bases long, suggesting that overlapping regions usually share a significant section of the two sets

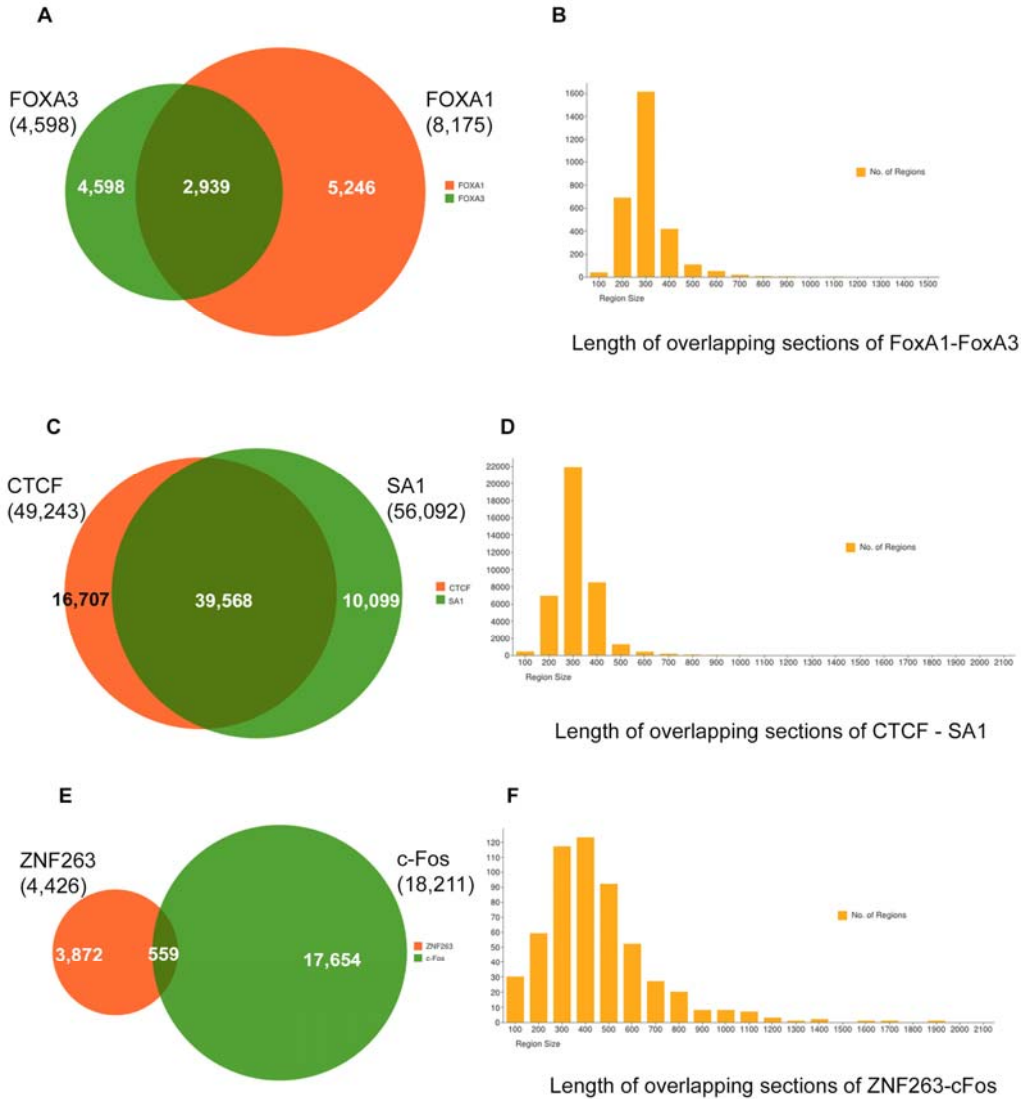


Figure 3-23: Example study of overlap between FOXA1 and FOXA3, CTCF and SA1, ZNF263 and c-Fos datasets. A) Venn diagram representation of 2,939 overlapping sections in FOXA1 (8,175 regions) and FOXA3 (4,598 regions) datasets. B) Histogram for bin size 100 showing size distribution of FOXA1-FOXA3 common sections of overlapping regions. C) Overlap between CTCF and SA1 datasets. D) Distribution of overlapping sections of CTCF and SA1. E) Overlap between ZNF263 and c-Fos datasets. D) Distribution of overlapping sections of ZNF263 and c-Fos.

We annotated the common sections of regions to observe their distribution and distance from the nearest TSS by setting criteria of 100kb up and downstream from TSS and extracted the annotations of the closest genes. BiSA reported 3656 gene annotations for FOXA1-FOXA3 overlapping sections, 45,508 annotations for CTCF-SA1 sections and 810 annotations for ZNF263-c-Fos sections. The annotation files also contain the distances from each TSS. Using these numbers we drew density plots for FOXA1-FOXA3 and CTCF-SA1 in R language showing the distribution of the overlapping regions upstream and downstream of TSSs (Figure 3.26). The plot reveals that the common regions of two datasets are concentrated around TSS suggesting the biological relevance of the overlap in agreement to the original publications.

Finally we investigated the statistical significance of overlap for each of the example comparisons. We calculated p-values for all regions in both datasets for each comparison using the hg18 whole genome domain. Selecting FOXA1 as query and FOXA3 as reference returned an overlap correlation value (OCV) of 0.50. By contrast, if FOXA3 was compared as query to FOXA1 as reference, the OCV was increased to 0.72. This provided an average OCV value of 0.61.

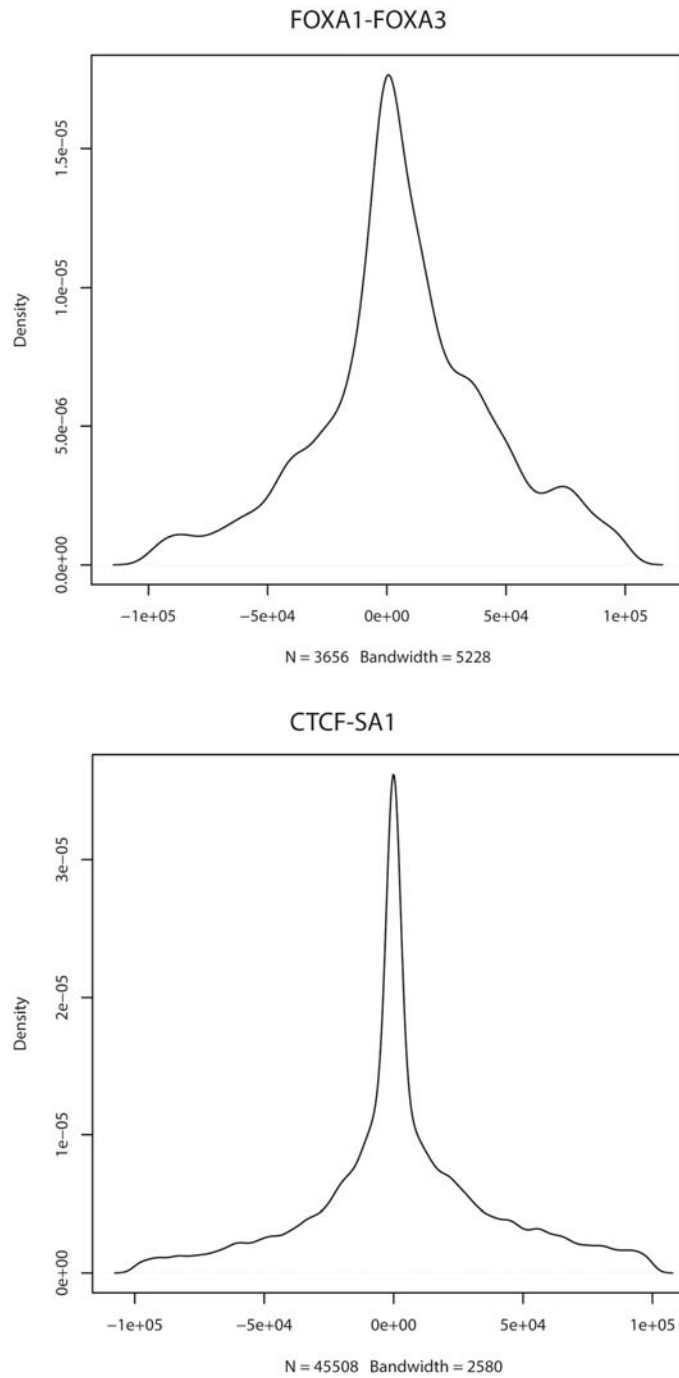


Figure 3-24: Density plot showing distribution of distance from TSS of genes. Common sections of regions from FOXA1-FOXA3 and CTCF-SA datasets were annotated, and drew density plots for the distribution of distances from the TSS.

An average OCV between FOXA1-FOXA3 above 0.5 suggests that two datasets significantly overlap, implying that the overlap between FOXA1 and FOXA3 is statistically significant. The higher OCV value (0.72) for FOXA3 when selected as query indicate that binding pattern of FOXA3 is highly collocating with FOXA1 binding while low OCV (0.5) for FOXA1 shows higher independence from the FOXA3 binding pattern. The high degree of overlap between CTCF and SA1 also returned a significant average OCV at 0.79. By contrast the lower level of overlap seen between ZNF263 and c-Fos was reflected in a non-significant average OCV of 0.19, confirming that the two TFs are not related and do not act on the same DNA regions in general.

3.6. Discussion

BiSA has been designed to meet the challenges of identifying genomic region overlaps in whole genome datasets. BiSA includes an up-to-date database of previously published studies reporting binding sites for different factors and specific histone modifications in a range of conditions and cell types. No tool, to our knowledge, includes such a pre-loaded knowledge base. Initially we have included data generated from human and mouse cells, and expansion to other organisms is easily possible. BiSA provides a user-friendly interface allowing the user to define and discover overlapping and nearby genomic regions either genome-wide or limited by chromosome. Users can visualize genomic overlap results as Venn diagrams and can save chart images for use in publications. BiSA can identify genes associated with binding regions of interest and also the statistical significance of overlapping regions.

Although the Apple Macintosh Unix and Linux environments are popular in genomic research, Windows-based informatics tools also exist (Ji et al., 2008; Khushi et al., 2012a; Khushi et al., 2012b). BiSA for Windows exploits the power of multi-core personal computers. In comparison to BiSA, most bioinformatics tools are command line, and such tools are not easy to install or to operate by the bench biologist. Galaxy (Goecks et al., 2010a)

offers a web-based tool 'Intersect', however it is limited in functionality. BiSA's Windows GUI is user-friendly for biologists and provides a sequential step-by-step guide through all the options. BiSA provides an easy interface to search and select KB based on organism, factor, cell line, condition, peak caller or first author name.

Most bioinformatics facilities run their servers on Linux/Unix or MAC operating systems, therefore, we have designed BiSA version for Linux/Mac that runs under Galaxy platform. Choosing the Galaxy platform saved us writing a lot of code and we have used Galaxy's built-in capability for managing users, security of datasets and sharing among other users. However, one disadvantage is that Galaxy has its own strict rules of what can be or cannot be done. One of the major issues that we observed with Galaxy was that programmatically it cannot be identified who is logged in until the user executes a tool. Therefore tools cannot be hidden from specific sets of users, however, when a user runs a tool then at its execution the user can be identified and restrictions can be applied.

Finally, a major strength of BiSA is the comprehensive knowledge base, coupled with tools to analyse overlapping regions, statistical significance of the overlapping regions and ability to annotate and visualize the regions of interest. BiSA's comprehensive KB is not only useful for rapid comparison of a user's own results to previously published datasets, but also to inform decisions such as selection of a peak caller programme or in comparing numbers of peaks. The KB suggests that MACS is a popular peak caller software in ChIP-Seq studies followed by Cisgenome and HOMER, whereas, the MAT algorithm is widely used in ChIP-chip studies. In summary, BiSA is designed for ease of use on a Windows platform, and includes a comprehensive knowledge base of binding site and histone modification datasets. BiSA has the potential to be a useful tool in identifying overlaps in genomic binding regions and histone modifications of common transcription factors for biologist.

Chapter 4: Significance of Transcription Factor Overlapping Regions

4.1. Introduction

Transcription factors (TFs) form complexes to regulate genes by either binding directly to specific DNA sequences or through recruitment of other DNA binding co-factors by a tethering mechanism (Wang et al., 2012). Therefore, if two TFs co-locate on the same genomic location or very close to each other then it is likely that the two factors form a complex or interact in some way rather than independently regulating gene expression. Targeting transcription factors via protein-protein interaction can offer a novel strategy for cancer therapy. For example, in many human cancers MDM2 binds to tumour suppressor transcription factor p53 and impairs p53 function. This led to the discovery of Nutlin, a small molecule inhibitor that perturbs the interaction between MDM2 and p53 (Johnston and Carroll, 2015; Vassilev et al., 2004; Klein and Vassilev, 2004). Therefore it is important in biology to identify interacting or partner TFs.

Another example of TF cooperation is in human liver hepatocellular cells (HepG2) where the analysis of FOXA1 and FOXA3 ChIP-Seq data identified a co-location of FOXA1, FOXA2, and FOXA3 suggesting that these FOXA family factors formed a complex. Further analysis using co-immunoprecipitation identified that FOXA2 interacts with FOXA1 and FOXA3 however, FOXA1 and FOXA3 did not interact (Motallebipour et al., 2009b).

In this chapter I demonstrate the utility of the BiSA knowledge base for generating biological hypotheses by studying statistical significance of transcription factor co-occurrence. I first perform a validation of statistical comparison among datasets that are generated using different tools. I generate a spreadsheet for quick observation of overlap among various factors. Finally, I model a cell-line specific transcription factor co-localisation network by

calculating statistical significance of overlap of transcription factor binding regions. Based on the bioinformatic analysis, this chapter generated some biological relevant hypotheses by exploring co-localisation of among various factors so that they could be further studied in the laboratory.

4.2. Methods

Transcription factors and histone marks act in a cell-specific manner (Arvey et al., 2012; Jolma et al., 2013; Eeckhoute et al., 2009). To study co-localisation of different factors the datasets should therefore be generated or derived from the same cell type. Secondly genomic coordinates significantly differ in different genomic assemblies, for example the transcription start site (TSS) for the BRCA1 gene in hg18 assembly is chr17:38449837 while the BRCA1 TSS in hg19 assembly is chr17:41196311. Therefore, for any comparison datasets must be from the same assembly.

The BiSA knowledge base (database) contains 1005 datasets of mouse and human assemblies consisting of ~24 million genomic regions. Figure 4.1 shows the distribution of datasets for four assemblies (hg19, hg18, mm9 and mm8) in the BiSA database. This distribution showed that most studies were conducted on human assemblies in order to understand the molecular mechanisms that drive diseases in human. There were significantly more same cell line datasets in hg19 than other assemblies therefore, in this chapter I analysed hg19 datasets.

The BiSA algorithm (RegMap) was used to identify regions that overlapped by at least 1 base pair. The analysis was primarily performed in BiSA for Windows with SQL Server due to the comparative better performance and ease of use. A script was written that generated the overlapping results for all datasets in the BiSA knowledge base (Box 4.1). The overlapping results were saved into a table (*OverlapAllStudy*) in the database. The results were exported

to a Microsoft Excel file and published (Khushi, 2015) for researchers who can browse and filter results without the need to install BiSA.

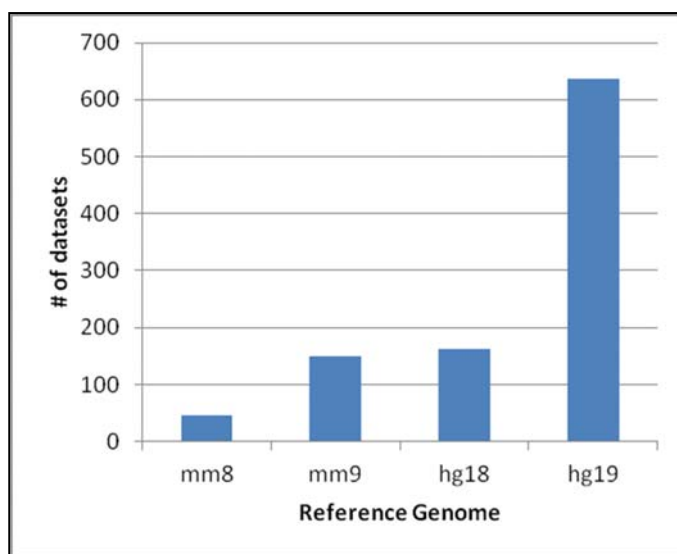


Figure 4-1: Distribution of datasets for four (mm8, mm9, hg18 and hg19) reference assemblies in the BiSA knowledge base.

In BiSA the IntervalStats tool (Chikina and Troyanskaya, 2012) is implemented to calculate the statistical significance of a factor overlapping with another factor as explained in Chapter 2. IntervalStats calculates a p-value for each peak region by comparing a region from a query dataset to the nearest region in a reference dataset. The tool restricts the analysis to regions that are within a domain dataset which can be a whole genome or can be possible interval locations such as promoter proximal regions. Based on IntervalStat calculated p-values, BiSA also calculates a summary statistic, that we refer to as the Overlap Correlation Value (OCV). The OCV ranges from 0 to 1, the closer the value to 1 the stronger the significance of overlap of two datasets. The OCV represents the fraction of regions in the query dataset with a p-value less than a specified threshold. For example for a threshold value of 0.05, if there were 60 regions in query dataset (total 100 regions) having p-value less than 0.05 then the OCV will be 0.6. For analysis in this chapter we set the p-value threshold to 0.05.


```

Declare @KBid_A int, @KBid_B int, @TotalOverlaps varchar(50)

DECLARE c_KB1 CURSOR FOR SELECT KBid from KBDetails where RefGenome='hg19'
OPEN c_KB1

    FETCH NEXT FROM c_KB1 INTO @KBid_A
    WHILE (@@FETCH_STATUS = 0) BEGIN
        DECLARE c_KB2 CURSOR FOR SELECT KBid from KBDetails where KBid <> @KBid_A
and RefGenome='hg19'
        OPEN c_KB2
        FETCH NEXT FROM c_KB2 INTO @KBid_B
        WHILE (@@FETCH_STATUS = 0) BEGIN
            if not exists (select * from OverlapAllStudy where (KBid_A=@KBid_A
and KBid_B=@KBid_B) or (KBid_B=@KBid_A and KBid_A=@KBid_B) ) BEGIN
                select @TotalOverlaps=count(*) from vwKBCompareSites where
KBid_A=@KBid_A and KBid_B=@KBid_B and bpOverlap>=1
                insert into OverlapAllStudy Values(@KBid_A,
@KBid_B,@TotalOverlaps)
            END

            FETCH NEXT FROM c_KB2 INTO @KBid_B
        End

        CLOSE c_KB2
        DEALLOCATE c_KB2

    FETCH NEXT FROM c_KB1 INTO @KBid_A
End
CLOSE c_KB1
DEALLOCATE c_KB1

```

Box 4.1: *SQL code to pair-wise count number of overlaps for datasets in human genome assembly, hg19.*

The ChIP-Seq analysis pipe-line involves identifying peak regions using peak-caller software as discussed in detail in Chapter 1. Different software identify different numbers of peaks and therefore, the total number of regions that overlap between two datasets can be subject to change and this could also change the OCV. According to the BiSA knowledge base the most popular peak caller used was MACS. There were ~82% (825/1005) datasets generated using MACS and many of the remaining studies had validated their peak-calling with MACS. The second most popular peak-caller was Homer (42 datasets). Therefore I investigated the variation in OCV when these two peak-callers were employed. I began by collecting raw sequence data from Welboren *et. al.* (Welboren et al., 2009). The authors performed ERα ChIP-Seq with three treatments: estradiol (E2), tamoxifen (Tam) and fulvestrant (Fulv). I

aligned raw sequences to the hg19 assembly using Bowtie 2 and called peaks using both MACS and Homer.

For calculating OCV, one dataset was selected as a query factor while the other was selected as a reference dataset; OCV can change when datasets are swapped between query and reference due to the different numbers and lengths of regions in each dataset. I translated this relationship into ‘directionality of a factor’. When OCV of a query factor was equal or greater than 0.5 then I considered the relationship to be significantly influenced by binding of the reference factor, I then draw this relationship with an arrow from the reference factor to the query factor. This method helped us in visualising significantly interacting co-factors and was employed to study datasets for a breast cancer cell line (T47D) which is one of the main studied disease in our research group.

4.3. Results

4.3.1. Validation of Overlap Correlation Value (OCV)

I performed a systematic validation to see how much the OCV varied if different peak-callers were used. Using the datasets from Welboren *et. al.* (Welboren et al., 2009) study, I called 6,172, 6,105 and 2,447 regions for ER α -E2, ER α -Tam and ER α -Fulv, respectively using MACS. On the other hand, using HOMER, I called 6,893, 6,320 and 2,430 regions for ER α -E2, ER α -Tam and ER α -Fulv, respectively. This indicated that the number of regions being called using the two peak-callers varied. Using BiSA, OCV was calculated for MACS datasets and then for HOMER datasets. I found there was a negligible difference in OCV when comparing the sets. For example, using MACS generated datasets when ER α -Fulv was selected as query and ER α -E2 was selected as reference the OCV was 0.67. The OCV decreased to 0.39 when datasets were swapped making average OCV 0.53 for ER α -E2 versus ER α -Fulv relationship. The average OCV for the same relation using HOMER datasets was

0.565. I considered this difference of 0.035 between the two average OCVs as insignificant. Similarly the differences for other comparisons were also minor (Figure 4.2).

I further investigated the variation in OCV by cross comparing MACS datasets against HOMER and vice versa. I again found either no or very small change in OCV (Table 4.1, Table 4.2). For example when ER α -E2 dataset, generated by MACS, were selected as query and ER α -Fulv dataset, generated by HOMER, were selected as reference the OCV was 0.38 (Table 4.1). The OCV remained unchanged (0.38) when ER α -E2 dataset, generated by HOMER, were selected as query and ER α -Fulv dataset, generated by MACS, were selected as reference (Table 4.2).

This validation identified that OCV remained either unchanged or there was a negligible change for both significant and non-significant interactions. Therefore I concluded that applying this statistical analysis on BiSA knowledge base is valid when datasets were generated using MACS or HOMER.

Query Datasets (MACS)	Reference Datasets (HOMER)		
	ER α -E2	ER α -Fulv	ER α -Tam
ERα-E2	1	0.38	0.55
ERα -Fulv	0.67	1	0.67
ERα-Tam	0.56	0.37	1

Table 4-1: OCV was calculated by selecting MACS datasets in first column (bold) as query while HOMER datasets were selected as reference.

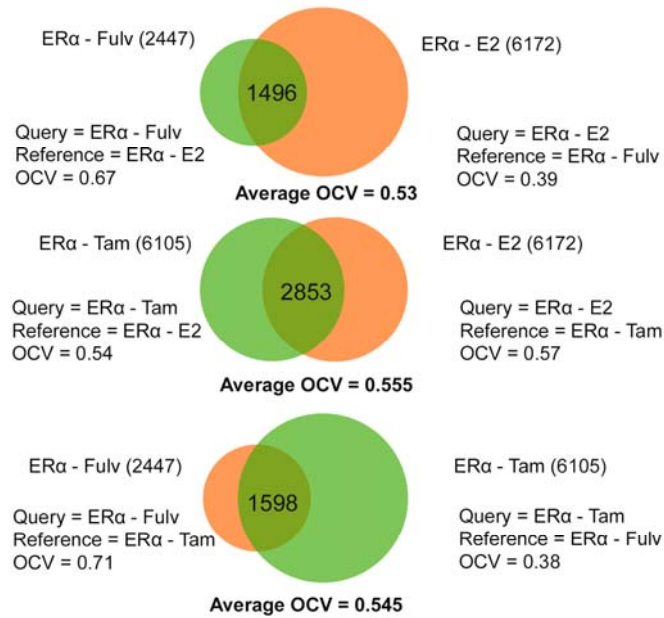
Query Datasets (HOMER)	MACS (Reference Datasets)		
	ER α -E2	ER α -Fulv	ER α -Tam
ERα-E2	1	0.38	0.56
ERα -Fulv	0.67	1	0.7
ERα-Tam	0.53	0.36	1

Table 4-2: OCV was calculated by selecting HOMER datasets in first column (bold) as query while MACS datasets were selected as reference.

4.3.2. Development of a Spreadsheet for Easy Identification of Degree of Overlap among Datasets

Using the BiSA overlap finding algorithm (RegMap) a Microsoft Excel file was generated containing information about cell line, factor name (either transcription factor or histone mark), treatment condition (if there is any), total number of regions, the number of overlapping regions and percentage found in other datasets. The results were spread across four spreadsheets depending on reference genome i.e hg19, hg18, mm9 or mm8 (Figure 4.3). Researchers can easily filter records based on restricting values in each field and then sorting on ‘Percentage Overlaps’ to find out the most or least interacting dataset. Links were also provided to the raw data and original publications.

Peak-calling with MACS



Peak-calling with HOMER

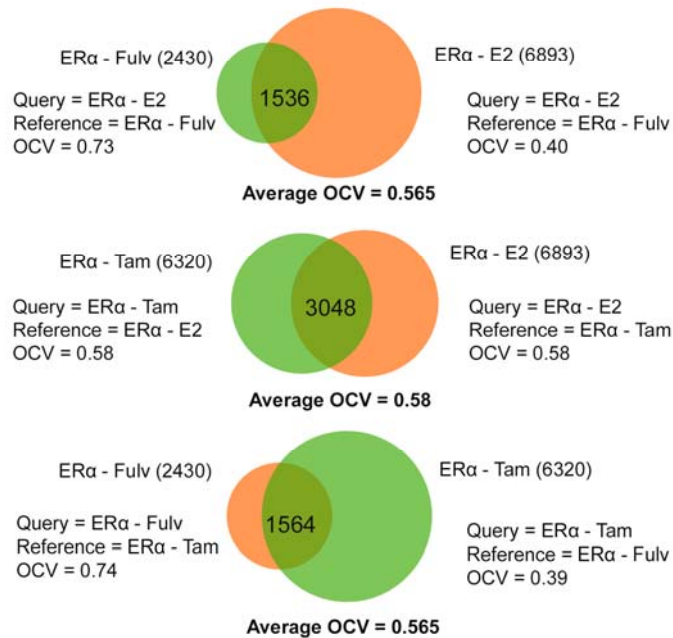


Figure 4-2: Peaks were called using MACS and HOMER tools for ERα datasets with E2, Tam and Fulv treatments. Pair-wise OCV was calculated to establish significance of variation. Venn diagrams show the degree of common regions among the two datasets.

Using the overlapping results data as described above, it was observed that in the HepG2 (liver hepatocellular) cell line, HNF4G (Gertz et al., 2013) preferentially bound (6,633/6,839, 97%) enhancer regions H3K4Me1 (Ram et al., 2011), and the majority of HNF4G binding sites (4,244/6,839, ~62%) were also found to overlap with STAG1 (Schmidt et al., 2010) binding sites (83,080 regions). 4,220 regions were common in the three datasets (Figure 4.4). The statistical significance of overlapping of HNF4G with STAG1 and H3K4me1 was further analysed in BiSA. BiSA revealed a statistically significant overlap correlation value (OCV=0.65) when HNF4G was selected as query and STAG1 was selected as reference dataset. Similarly OCV for HNF4G against H4K4me1 was also significant (OCV=0.5) (Figure 4.4).

HNF4G is an orphan nuclear receptor whose ligand and function has not been fully understood, however recent studies have shown HNF4G overexpression induces growth of cancer tissue (Yang et al., 2014; Okegawa et al., 2013). On the other hand, STAG1 (Stromal Antigen 1), also known as SA1, is one of the four subunits of the cohesin complex (Losada, 2014). Cohesin has important roles in transcription regulation, DNA repair, chromosome condensation, homolog pairing, etc. (Mehta et al., 2013; Losada, 2014). Therefore, statistically significant overlap of HNF4G with STAG1 indicates an important underlying biology which could be further explored in the laboratory. These results along with the Excel file containing overlapping results were published in my database benchmarking article (Khushi, 2015).

Dataset-A										Dataset-B							
Pub	RefLink	RawDataLink	CellLine	Condition	Factor	TotalReqs	overlaps	Percent	TotalRegions2	Factor2	Condition2	CellLine2	RawDataLink2	RefLink2	PubYear		
2013	http://www	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	6800	99.4%	56226	HNF4A	No treatment	HepG2	(live http://cistrome.oi	http://www.nc	2010		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	6633	97.0%	138709	H3K4me1		HepG2	(live http://genome.uc	http://www.nc	2011		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	6542	95.7%	19429	Hnf4a		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	6441	94.2%	32849	HNF4A	No treatment	HepG2	(live http://cistrome.oi	http://www.nc	2009		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	5143	75.2%	48936	H3k4me3		HepG2	(live http://genome.uc	http://www.nc	2011		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	4244	62.1%	83080	STAG1	No treatment	HepG2	(live http://cistrome.oi	http://www.nc	2010		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	4148	60.7%	34651	Foxa1		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	4142	60.6%	14958	P300		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	4089	59.8%	19191	Mybl2		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	4012	58.7%	119109	CEBPa	No treatment	HepG2	(live http://cistrome.oi	http://www.nc	2010		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	3727	54.5%	26493	Foxa1		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	3617	52.9%	13908	Nfic		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	2980	43.6%	57071	RAD21	No treatment	HepG2	(live http://cistrome.oi	http://www.nc	2010		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	2903	42.4%	7274	Hdac2		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	2482	36.3%	25385	Hey1		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	2324	34.0%	82376	H3k27me3		HepG2	(live http://genome.uc	http://www.nc	2011		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	2144	31.3%	8040	Mbd4		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	2055	30.0%	31151	Pol2		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	1992	29.1%	10237	Foxa2		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	1990	29.1%	10588	Cebp		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	1655	24.2%	19079	Jund		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	1554	22.7%	42288	Rad21		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	1527	22.3%	24194	Pol2		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	1334	19.5%	14232	Eif1		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	1219	17.8%	73607	CTCF	No treatment	HepG2	(live http://cistrome.oi	http://www.nc	2010		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	1184	17.3%	15848	Fosl2		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	1060	15.5%	7329	Zbtb7a		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	1048	15.3%	7100	Cebp		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	1046	15.3%	8148	Rxra		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	1008	14.7%	2926	Nr2f2		HepG2	(live http://www.ncbi.	http://www.nc	2013		
2013	http://ww	http://www.nd	HepG2	(liver hepatocellular	Hnf4g	6839	987	14.4%	3120	Tead4		HepG2	(live http://www.ncbi.	http://www.nc	2013		

Figure 4-3: Spread sheet for easy identification of degree of overlap among different datasets. Datasets are segregated based on their reference assembly, records can be filtered by clicking on small arrow heads on the top of columns.

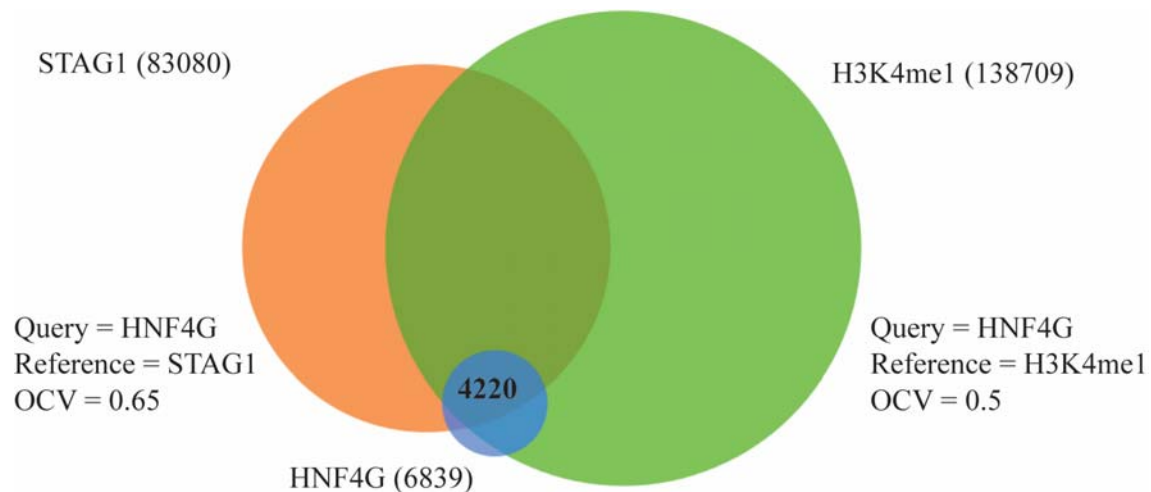


Figure 4-4: Venn diagram showing degree of overlap among HNF4G, STAG1 and H3K4me1 datasets.

4.3.3. Directionality of Transcription Cooperation

I studied degree of co-localisation of two factors by calculating OCV. As described in the methods section, when OCV of a query factor was equal or greater than 0.5 I considered this

to be binding of the query factor depended on the reference factor and was illustrated this by drawing an arrow from the reference factor pointing to the query factor. For example, when JUND was selected as a query factor against P300 reference, the OCV was significant (0.78), however, when P300 was selected as query against JUND reference then OCV was non-significant (0.37). This relationship was drawn as a one-way arrow pointing to JUND (Figure 4.5 A-ii). A two way arrow was used to represent two-way co-localisation. For example, there was significant two-way co-localisation of PR binding sites when treated with R5020 or ORG2058 (Figure 4.5 B). By visualising directionality of transcription factor interactions we could easily identify the potential co-factors that justify further study of the factors.

4.3.4. Transcription Factor Networks in ER/PR Positive Breast Cancer

Our research group study breast cancer which is one of the leading causes of cancer related deaths in the world (Kanavos, 2006; Jemal et al., 2011). The BiSA knowledge base contains a number of datasets for the T47D breast cancer cell line which is an ER α and PR positive breast cancer cell line. These steroid hormone receptors play a critical role in development and progression of breast cancer, therefore to find novel interacting partners I studied the statistical significance of co-location of different factors with ER α and PR in T47D cells.

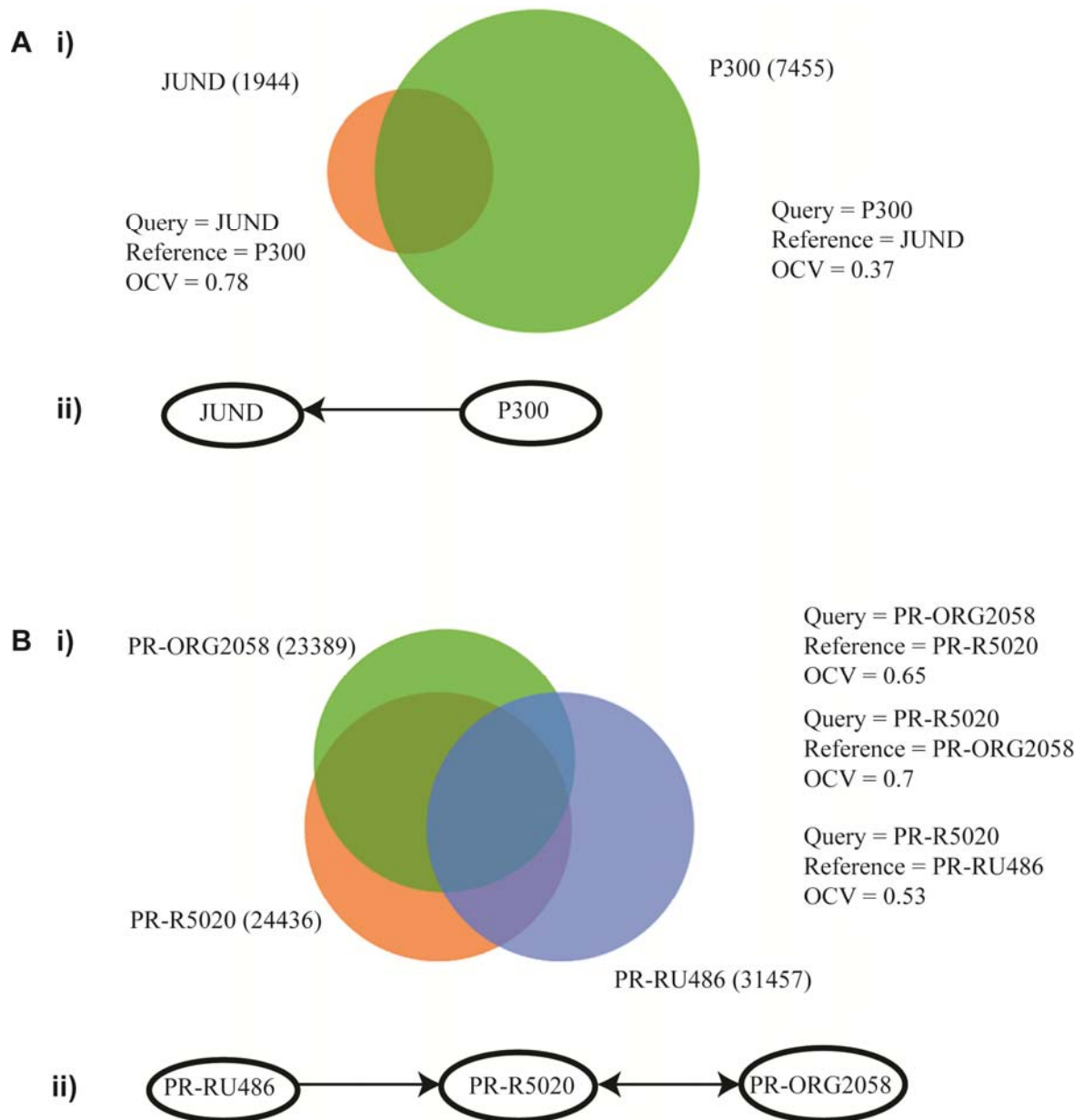


Figure 4-5: The degree of transcription factor overlap was transcribed into a network diagram. A-i) JUND was selected as a query factor against P300 reference, the OCV was significant (0.78), however, when P300 was selected as query against JUND reference then OCV was non-significant (0.37). The JUND-P300 relationship is shown as one-way arrow pointing to JUND (A-ii). B-i) Overlap between three PR datasets generated by three treatments. Out of six possible combination, three relationships having significant OCV

(>0.5) are shown. B-ii) The relationship between PR-R5020 and PR-ORG2058 was significant when either factor was selected as a query factor, shown by two-way arrow.

There were a number of ER α and PR datasets for T47D available in the BiSA knowledge base, therefore, I investigated which dataset would be most appropriate to study against other factors. For example, Yin et al. (Yin et al., 2012) generated PR binding sites by treatment with anti-progestin RU486 (mifepristone) and reported 31,457 binding regions. Whereas, to study PR binding regions Ballare *et al* (Ballare et al., 2013) treated T47D cells with 10 nM R5020 for different lengths of time, Clarke and Graham (Clarke and Graham, 2012) treated samples with 10 nM ORG2058 for 45 minutes before performing ChIP. I performed OCV statistical testing on how these datasets compare to each other in order to choose datasets for subsequent data analysis. Table 4.3 shows the OCV among various PR datasets.

Query datasets	Reference datasets				Average OCV for R5020 treatment		
	PR-RU486 60 min	PR- ORG2058 45 min	PR-R5020 5 min	PR-R5020 30 min		PR-R5020 60 min	PR-R5020 360 min
PR-RU486	X	0.33	0.39	0.39	0.37	0.34	0.65 0.72 0.74 0.70
PR-ORG2058 45min	0.44	X	0.48	0.64	0.64	0.64	
PR-R5020 5min	0.61	0.56	X	0.7	0.66	0.6	
PR-R5020 30min	0.55	0.67	0.64	X	0.79	0.72	
PR-R5020 60min	0.53	0.7	0.63	0.82	X	0.77	
PR-R5020 360min	0.49	0.69	0.57	0.76	0.78	X	

Table 4-3: Overlap Correlation Value (OCV) for PR datasets with various treatments. OCV was calculated by selecting datasets from first column (bold) as query and datasets from other column datasets as reference. Ligand concentrations were 10 nanometre (nM).

I identified that the PR activation and binding due to treatment of ORG2058 or R5020 was very similar and the correlation among these datasets were statistically significant (Table 4.3). The OCVs for the PR dataset treated with ORG2058 for 45 minutes against PR-R5020 datasets were more than 0.5. ORG2058 and R5020, being agonists to PR, generally exhibit very similar properties, however, to identify any difference in the datasets I decided to study both datasets against other factors in my further analysis.

The correlation among PR datasets, generated with treatment of R5020 at 5, 30, 60 and 360 minutes, was also significant, however, I identified that the PR dataset treated with R5020 at 60 minutes was most representative of the other 3 datasets, with average OCV 0.74 (Table 4.3). Therefore I decided to use this dataset in further analysis and labelled this dataset as PR-R5020.

On the other hand PR activation and resultant binding due to anti-progestin (RU486) treatment revealed non-significant OCVs against PR datasets that were generated by progestin activation which was in agreement with other studies that a factor's binding pattern was different using antagonist treatment compared to agonist treatment (Yin et al., 2012). Therefore I labelled the PR dataset treated with RU486 as PR-RU486 in my subsequent analysis.

Similarly, I had five datasets describing ER α binding sites from two studies (Joseph et al., 2010; Gertz et al., 2012). Gertz *et al* published ER α binding sites by treating with 6 μ L of 5000 \times concentrated E2 (Estradiol), GEN (Genistein) and BPA (Bisphenol A), while Joseph *et al.* treated with ethanol and E2. I calculated OCVs for these datasets (Table 4.4) against each other. I did not include ER α datasets generated after ethanol treatment for my subsequent analysis, since ER α is a ligand activated nuclear receptor. A higher OCV for ER α -BPA and ER α -GEN query datasets using the ER α -E2 dataset as reference revealed that

most ER α -BPA and ER α -GEN binding regions overlapped with E2 treatment dataset (Table 4.4). This confirmed that BPA and GEN treatment regulated only a subset of ER α -E2 regulated transcripts as found by the original study, therefore, the ER α -BPA and ER α -GEN datasets were not considered further. There was another ER α -E2 dataset from the Joseph *et al.* study (Joseph et al., 2010), however, I finally chose the Gertz *et al.* ER α -E2 dataset because the study reported more reads and higher number of binding regions (7,587 as compared to 5,437). In addition, JUND, CTCF and P300 binding datasets were also taken from the Gertz *et al.* study (Gertz et al., 2013), however, FOXA1 datasets were taken from Joseph *et al.* study (Joseph et al., 2010). GATA, JARID1B and XBP1 binding datasets were taken from Adoma *et al.*, Yamamoto *et al.* and Chen *et al.* studies respectively (Adomas et al., 2014; Yamamoto et al., 2014; Chen et al., 2014). 12 datasets for the T47D cell line were selected to study statistical significance of their co-localisation with each other (Table 4.5) and a heat map was drawn using R package gplots with hierarchical clustering (Figure 4.6). A network of transcription factor colocation (Figure 4.7) was also drawn using Dia tool (dia-installer.de).

Query Datasets	ER α -EtOH	ER α -BPA	ER α -GEN	ER α -E2 (Joseph et. al.)	ER α -E2 (Gertz et. al.)
ERα-EtOH	1	0.18	0.19	0.34	0.19
ERα-BPA	0.35	1	0.97	0.84	0.99
ERα-GEN	0.25	0.62	1	0.74	0.99
ERα-E2 (Joseph et. al.)	0.29	0.44	0.6	1	0.69
ERα-E2 (Gertz et. al.)	0.2	0.49	0.73	0.63	1

Table 4-4: OCVs among different ER α datasets. OCV is calculated by selecting datasets in the first column (bold) as query and datasets from other columns were selected as reference.

Joseph et al treated samples with 10 nM E2 for 3 hour while Gertz et al treated samples with 6 μ L of 5000 \times concentrated ligands for 10 min.

	Reference Datasets											
Query Datasets	FOXA1-DMSO	FOXA1-E2	PR-R5020	PR-ORG2058	PR-RU486	JUND	CTCF	P300	ER α -E2	GATA3	JARID1B	XBP1
FOXA1-DMSO	1	0.37	0.25	0.23	0.47	0.16	0.13	0.27	0.2	0.25	0.31	0.16
FOXA1-E2	0.57	1	0.23	0.22	0.39	0.14	0.14	0.27	0.23	0.24	0.36	0.14
PR-R5020	0.42	0.26	1	0.7	0.53	0.15	0.1	0.25	0.22	0.25	0.21	0.19
PR-ORG2058	0.34	0.22	0.64	1	0.44	0.14	0.09	0.21	0.21	0.22	0.17	0.17
PR-RU486	0.65	0.32	0.37	0.33	1	0.15	0.11	0.25	0.2	0.25	0.28	0.16
JUND	0.96	0.67	0.43	0.45	0.86	1	0.05	0.78	0.56	0.74	0.28	0.37
CTCF	0.26	0.17	0.11	0.1	0.18	0.06	1	0.1	0.08	0.11	0.75	0.1
P300	0.94	0.61	0.43	0.4	0.8	0.37	0.11	1	0.45	0.52	0.5	0.32
ERα-E2	0.65	0.57	0.39	0.41	0.58	0.29	0.08	0.47	1	0.5	0.29	0.28
GATA3	0.63	0.39	0.32	0.32	0.55	0.25	0.11	0.36	0.33	1	0.28	0.23
JARID1B	0.44	0.32	0.18	0.16	0.3	0.09	0.35	0.21	0.14	0.17	1	0.12
XBP1	0.6	0.35	0.36	0.35	0.55	0.22	0.12	0.37	0.28	0.34	0.36	1

Table 4-5: OCVs among 12 compared datasets. Datasets in the first column (bold) were select as query and the datasets from columns were selected as reference in the calculations of OCV. OCVs ≥ 0.5 are shown in red.

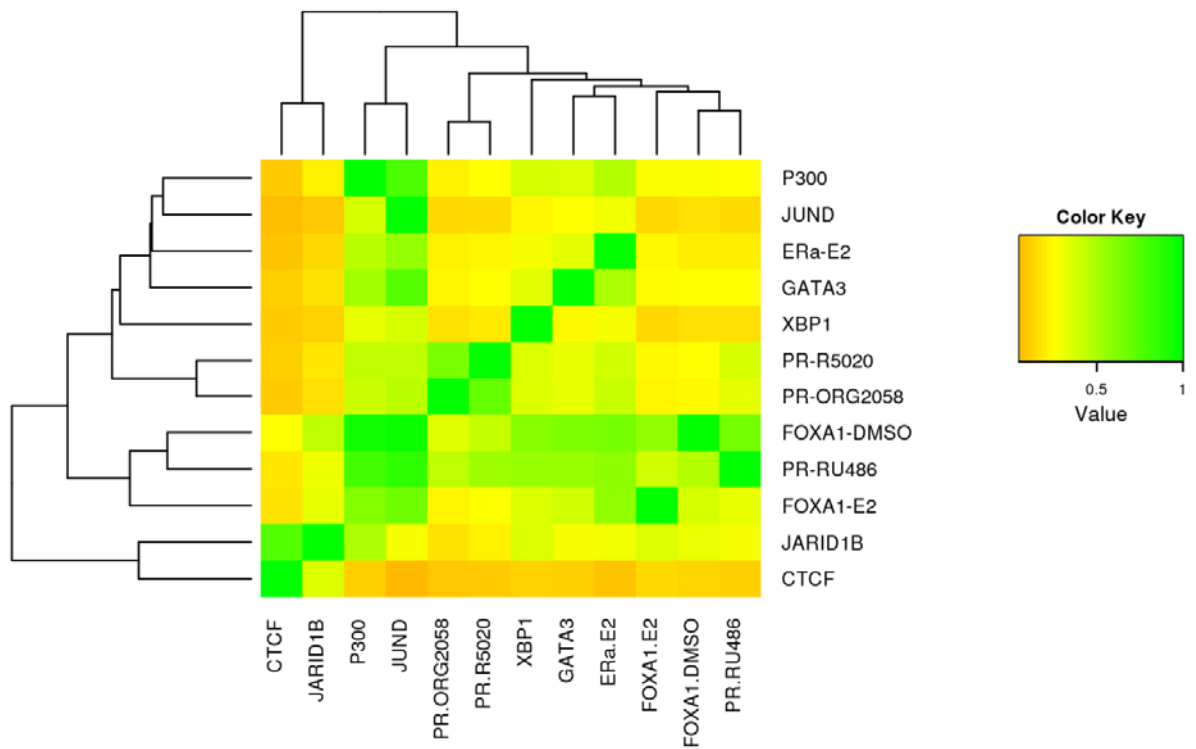


Figure 4-6: Hierarchical clustering heat map showing correlation of 12 datasets in T47D cells using OCV calculated in Table 4.5.

ER α binding stimulated by E2 revealed a significant correlation (OCV=0.58) with PR binding stimulated with RU486 (anti-progestin) in comparison to PR binding stimulated by progestin treatment (OCV=0.33). ER α binding also showed a significant correlation with FOXA1 and GATA3 when ER α was selected as query factor. On the other hand, ER α binding influenced JUND binding when JUND was selected as a query factor (OCV=0.56). JUND query also revealed a significant correlation with ER α , P300, PR-RU486, FOXA1 and GATA3 (Figure 4.7).

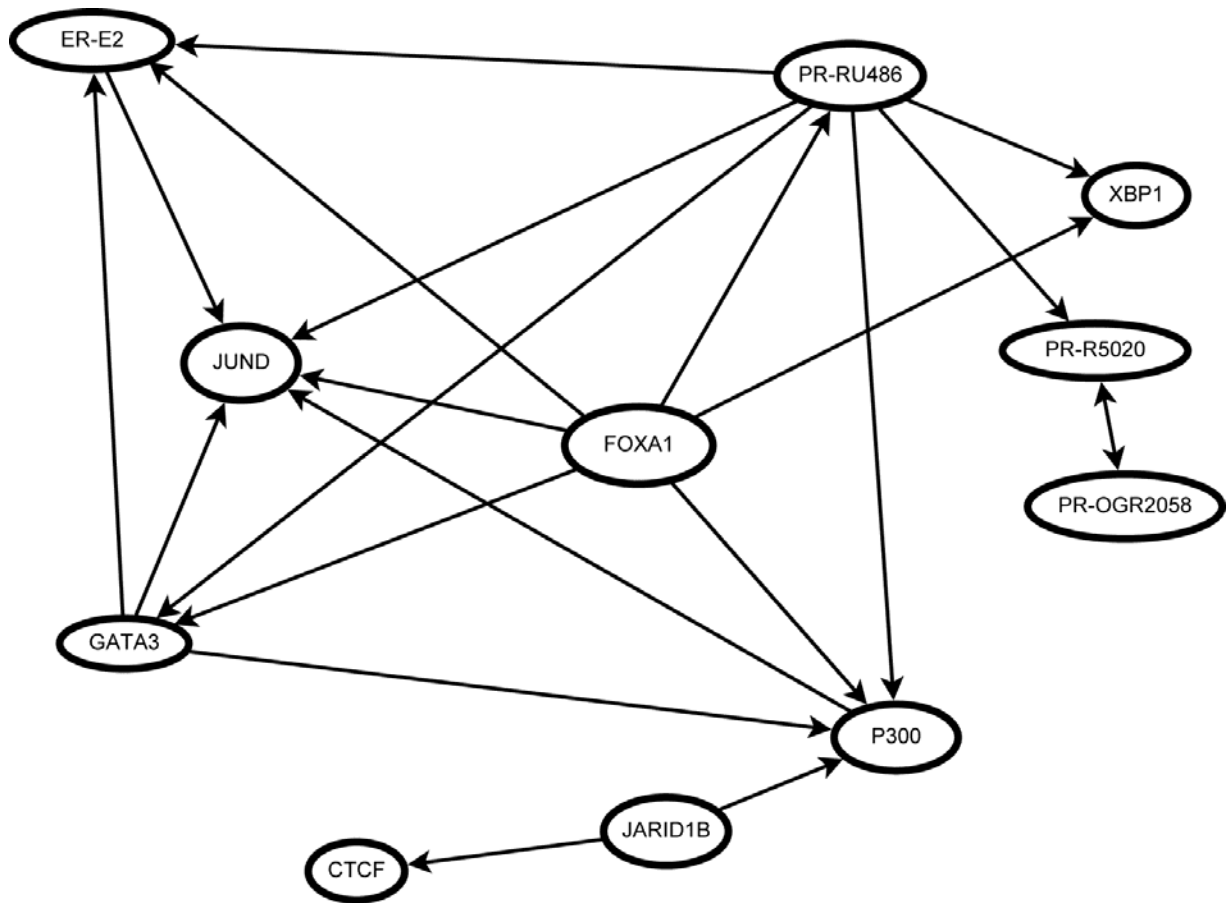


Figure 4-7: Transcription factor network in T47D breast cancer cell line. . Arrow head points towards a query factor whose OCV was greater than 0.5.

When the OCV of a query factor was greater than 0.5 an arrow was drawn pointing to query from reference factor.

When P300 was selected as a query factor it showed a significant correlation with PR-RU486, FOXA1 and GATA3. P300 is a transcription co-activator which plays a critical role in cell growth, proliferation, oncogenesis, apoptosis progression and development of diseases (Ghosh and Varga, 2007; Shikama et al., 2003; Partanen et al., 1999; Zhang et al., 2014).

Unlike transcription factors P300 does not bind directly to DNA, the molecule has a number of structural domains such as histone acetyltransferase (HAT) domain, and a C-terminal

glutamine-rich domain (Xu et al., 2001) which enables it to interact with nuclear receptors such as ER α /PR and other transcription factors such as GATA3, JARID1B and JUND (Figure 4.4). The P300 dataset demonstrated strong correlation with PR-RU486 OCV (0.8) and was almost twice the value reported by PR-R5020 (OCV=0.43) or PR-ORG2058 (OCV=0.4). P300 OCV with ER α -E2 was also low (OCV 0.45).

FOXA1 is a pioneer factor that facilitates the binding of ER α and other factors (Carroll et al., 2005; Carroll et al., 2006; Hu et al., 2014; Jin et al., 2014). FOXA1 datasets when selected as the query against all other factors the OCV was not significant (Table 4.5, Figure 4.7). This statistical analysis confirmed previous findings that FOXA1 was a pioneer factor and its binding was independent of binding of other factors. On the other hand, co-location of JUND, P300 and ER α with FOXA1 datasets (DMSO and E2 treatment) revealed a significant OCV. Co-location of XBP1, GATA3 and PR-RU488 only revealed a significant co-location with FOXA1-DMSO dataset. Therefore in summarising these relationships in Figure 4.7, I used the FOXA1- DMSO dataset to show the relationship with other factors that significantly facilitates binding of JUND, P300, ER α , GATA3 and PR-RU486, XBP1.

CTCF is a silencing factor and its co-location with cohesin components SA1 and RAD21 is known in some cell-types (Lee and Iyer, 2012; Herold et al., 2012; Fiorentino and Giordano, 2012; Rubio et al., 2008). CTCF binding revealed no significant overlap with other factors except JARID1B. Therefore I concluded that CTCF form homodimers by binding to itself, therefore, its binding to DNA results in tightly bound chromatin where these regions become unavailable for binding of other factors (Yusufzai et al., 2004; Holwerda and de Laat, 2013). JARID1B also known as PLU-1 was found highly expressed in some cancers including breast cancer (Yamane et al., 2007), therefore, the significant co-location (OCV=0.75) of CTCF with JARID1B identified an interesting biological correlation which should further be analysed.

4.4. Discussion

BiSA has a comprehensive knowledge base. By analysing a large number of datasets in a systematic way we can unveil interesting systems biology or generate new hypotheses that can further be explored in the laboratory.

BiSA employs Structured Query Language (SQL) which is a powerful query language. Once a proper table structure was setup, it took only 12 seconds to count the number of regions on each chromosome in the hg19 genome assembly while there were total of ~24 million genomic regions in the BiSA database. Easy SQL statement can be used to query the whole database of unlimited number of datasets, while equivalent searching tasks require writing lengthy code in other languages. However installation of Microsoft SQL Server could be a non-trivial task for many biologists, therefore, I published an easy navigable spread sheet file for identification of interesting overlap among datasets (Khushi 2015). This Excel file provides filters on the fields that provision limiting records based on cell line, factor or percentage overlap. Links to original publications and raw data are also provided in the file. Therefore, once an interesting overlap is identified a researcher can study the actual regions in BiSA or download original data for studying in other tools.

The analyses presented in this chapter can be performed without the BiSA by developing a command-line tool in a conventional language such as Python and the performance could be compared against the database-driven BiSA. However, developing any such tool is outside the scope of this thesis.

Statistical significance of degree of overlap between two datasets was calculated using BiSA embedded IntervalStats which calculates p-value of each query region against overlapping or closest reference region. Implementing this statistical technique on large datasets is very slow. IntervalStats is not a multi-threaded application, and consequently cannot take

advantage of multiple computational cores. My validation shows that this technique can be applied to datasets that were generated using the two most popular peak-caller tools MACS and Homer. By calculating the OCV, I studied the overlap of 12 datasets of various factors under different treatment conditions and then generated a network diagram. The network map of these factors revealed interesting biological interactions among various factors. The network map revealed that FOXA1 influenced binding of seven other factors JUND, P300, ER α , GATA3 and PR-RU486, XBP1, however, its own binding was independent of other factors. Interestingly ER α had significant OCV with PR when stimulated by anti-progesterone (RU-486) while ER α showed no significant overlap with PR when simulated with progesterone (R5020 or ORG2058).

Mifepristone (RU-486), abbreviated as MFP, is one of selective progesterone receptor modulators as this synthetic compound binds to PR and exhibits phenotypes ranging from agonism and antagonism (Benagiano et al., 2008). My analysis showed that the binding sites that were targeted by PR were different under progesterone and mifepristone treatment, however, there were 8801 regions common regions between PR-RU486 and PR-R5020 which supports previous results that mifepristone acts as partial agonist in the absence of progesterone. Mifepristone is used as an abortifacient in first trimester, emergency contraception and in a low dose as contraceptive medicine (Benagiano et al., 2014).

Tristan *et al* 2012 reviewed the literature for MFP effect on uterine fibroids and concluded that MFP reduced heavy menstrual bleeding and improved quality of life with no effect on fibroid volume (Tristan et al., 2012). However Yerushalimi *et al* showed that MFP vaginal treatment of 10mg/day significantly reduced the volume of fibroids from $135.3 \pm 22.9 \text{ cm}^3$ to $101.2 \pm 22.4 \text{ cm}^3$ after 3 months of treatment (Yerushalmi et al., 2014). Reduction in fibroid size with MFP treatment suggests that progesterone is the primary stimulant in uterine leiomyoma instead of estadiol (Benagiano et al., 2014; Chabbert-Buffet et al., 2014).

Therefore the ER α -E2 significant overlap (OCV=0.65) with PR-RU486 identify an important biological cooperation among the two factors.

CTCF, a silencing factor, showed a significant colocation with JARID1B (OCV=0.75).

Recently it has been shown that deletion of the JARID1B gene induces apoptosis in two cell lines for mantle cell lymphoma and acute myeloid leukemia, confirms an important role of JARID1B in carcinogenesis (Su et al., 2015). In another study it has been shown that up-regulation of JARID1B was related to poor prognosis and chemotherapy resistance in ovarian cancer (Wang et al., 2015). Furthermore its role in breast cancer and other cancers is also known (Yamane et al., 2007; Xiang et al., 2007; Yamamoto et al., 2014), therefore, statistically significant binding of CTCF with JARID1B identified an important biological correlation which should further be explored.

In summary this chapter has shown the usefulness of the BiSA knowledge base. Using data from three different studies I identified that the HNF4G nuclear receptor significantly collocates with STAG1 and H3K4me3 promoter marks in the HepG2 cell line. Finally I drew a transcription factor action network in the T47D cell-line by extracting data from various studies. This network map revealed that PR binding as a result of RU486 (anti-progestin) treatment significant more active and co-locate with many other factors than previous thought.

Chapter 5: Bioinformatic analysis of cis-regulatory interactions between progesterone and estrogen receptors in breast cancer

5.1. Introduction

The ovarian steroid hormones progesterone and estrogen play critical roles in the development and progression of breast cancer and endometriosis (Salehnia and Zavareh, 2013; Shao et al., 2014; D'Abreo and Hindenburg, 2013). These hormones exert their functions by activating specific nuclear receptors, estrogen binds to estrogen receptor ($ER\alpha$) and progesterone binds to progesterone receptor (PR) (Tsai and O'Malley, 1994).

Once activated these receptors bind to their DNA response elements and regulate transcription of target genes. $ER\alpha$ and PR, along with human epidermal growth factor receptor 2 (HER2), are used to classify phenotypes in breast cancers and to predict response to specific therapies (Kittler et al., 2013; Cadoo et al., 2013). A high number of $ER\alpha$ positive breast cancers are also PR positive (Cadoo et al., 2013; Penault-Llorca and Viale, 2012). Furthermore, studies from animal models and clinical trials have shown that progesterone via its receptor PR is a major player in development and growth of breast cancer and uterine fibroids, however, PR inhibits the development of estrogen-driven endometrial cancer (Kim et al., 2013a; Ishikawa et al., 2010). Many recent reviews highlight the importance of the role that progesterone and estrogen play via their receptors in various types of breast cancers (Yadav et al., 2014; Abdel-Hafiz and Horwitz, 2014; Obiorah et al., 2014; Kalkman et al., 2014; Wang and Di, 2014). Therefore it is important to understand how $ER\alpha$ and PR work together in regulating a number of cellular pathways, and clinical and molecular research on these factors continue to unveil new insights (Bulun, 2014).

It is acknowledged that $ER\alpha$ and PR binding, as well as that of other steroid hormone receptors, is assisted by binding of the pioneer transcription factor FOXA1 (Ballare et al.,

2013; Lam et al., 2013) to condensed chromatin, therefore, the interactions of FOXA1 with other factors have been well studied (Bernardo and Keri, 2012; Augello et al., 2011). There are a number of publications that have studied PR binding sites in progesterone-treated breast and other tissues (Yin et al., 2012; Clarke and Graham, 2012; Ballare et al., 2013). Many studies have also published ER α binding sites (Tsai et al., 2010; Schmidt et al., 2010; Joseph et al., 2010). However there is lack of investigation into the combined action of the two factors on DNA. Therefore in this chapter we investigated the interaction of these nuclear receptors on DNA. Our previously described (Chapter 3) and published BiSA database (Khushi et al., 2014) contains a number of datasets describing ER α and PR binding sites for various cell lines, therefore, we investigated the binding pattern of these factors in the T-47D breast cancer cell line. T-47D cells are derived from metastatic female human breast cancer and are known to be ER α and PR positive and their growth is simulated by treatment with estrogen (Ström et al., 2004; Chalbos et al., 1982).

5.2. Methods

PR data were taken from the study of Clarke and Graham (Clarke and Graham, 2012) and ER α data were obtained from the ENCODE project (Gertz et al., 2012). PR data were obtained by treating T47D cells with the progestin ORG2058 for 45 minutes, followed by PR-specific chromatin immunoprecipitation and deep sequencing (ChIP-Seq). Gertz et al studied ER α binding sites by treating with estradiol (E2), GEN (Genistein) and BPA (Bisphenol A) and conclude that compared to E2, GEN and BPA treatment results in fewer ER α binding sites and less change in gene expression. We selected the E2-treated dataset for our study. Datasets from both studies were of 36 base pair lengths generated on the Illumina platform. The PR data were generated using an Illumina Genome Analyzer Iix while ER α libraries were sequenced on Illumina HiSeq 2000. The data used in this study have been derived from peer-reviewed publications, suggesting that they are of an acceptable quality, in

addition we also performed standard quality control checks prior to our re-analysis of the raw data. The two studies used different genome assemblies and different tools to align the reads and to call the peaks. Therefore, to remove any biases we re-analysed the raw ER α and PR data. We mapped the raw data to the GRCh37/hg19 human genome assembly using Bowtie version 2 (Langmead and Salzberg, 2012). The aligned replicates were merged using Picard tools (Li et al., 2009) and the Model-based Analysis of ChIP-seq algorithm (MACS) version 1.4.2 (Zhang et al., 2008) was employed, with default settings, to identify PR and ER α binding regions in the two datasets. Regions associated with greater than 5% false discovery rate (FDR) were removed (Zhang et al., 2008).

We performed motif analysis using HOMER software (Heinz et al., 2010). HOMER employs a differential motif discovery algorithm by comparing two sets of sequences and quantifying consensus motifs that are differentially enriched in a set. HOMER automatically generates an appropriate background sequence matched for the GC content to avoid bias from CpG Islands. The tool is exclusively written for analysing DNA regulatory elements in ChIP-Seq experiments and has been used in number of high impact publications (Berman et al., 2012; Wang et al., 2011b; Xie et al., 2013).

Overlapping features were studied in BiSA (Khushi et al., 2014). BiSA is a bioinformatics database resource that can be run on Windows as a personal resource or web-based under Galaxy (Goecks et al., 2010a) as a collaborative tool. BiSA is pre-populated with published transcription factor and histone modification datasets and allows investigators to run a number of overlapping and non-overlapping genomic region analyses using their own datasets, or against the pre-loaded Knowledge Base. Overlapping features can be visualised as a Venn diagram and binding regions of interest can also be annotated with nearby genes. BiSA also provides an easy graphical interface to find the statistical significance of observed overlap between two genomic region datasets by implementing the IntervalStats tool

(Chikina and Troyanskaya, 2012). The tool calculates a p-value for each peak region by comparing a region from the query dataset to all regions in a reference dataset. The tool restricts the analysis to regions that are within a domain dataset which can be a whole genome or can be possible interval locations such as promoter proximal regions. Based on IntervalStat calculated p-values BiSA calculates a summary statistic, that we refer to as the Overlap Correlation Value (OCV). The OCV ranges from 0 to 1, the closer the value to 1 the stronger the significance of overlap of two datasets. The OCV represents the fraction of regions in the query dataset with a p-value less than a specified threshold. In BiSA, we have set the threshold p-value to 0.05 and used a number of domains such as whole genome and promoter proximal regions for this analysis.

We also investigated the spatial correlation of regions of whole datasets being closer to each other by Binary Interval Search (BITS) (Layer et al., 2013) and Genometricorr (Favorov et al., 2012). BITS implements a Monte Carlo simulation by comparing actual overlapping regions to random observed overlap. Genometricorr considers one genomic region set as a reference and the other set as a query and provides four asymmetric pair-wise statistical tests i) relative distance also called Local Correlation, ii) Absolute Distance, iii) Jaccard statistic and iv) Projection statistical tests. In Local Correlation the significance of relative distance between the genomic regions is measured by Kolmogorov-Smirnov test; in absolute distance test the significance of base pair distance among the regions is measured by permutation test; the Jaccard statistic takes into account the ratio of intersecting bases to the union base pairs. A Projection test calculates the overlapping centre points of query to reference regions and finds the significance of a result outside of the null expectation by binomial test (Favorov et al., 2012). We performed 10,000 simulations for BITS and Genometricorr statistical tests.

We performed functional annotation of ER α -PR common cis-regulatory regions using GREAT (Genomic Regions Enrichment of Annotations Tool) (McLean et al., 2010). GREAT

incorporates annotations from 20 ontologies covering gene ontology, phenotype data, human disease pathways, gene expression, regulatory motifs and gene families. We performed GREAT annotation using its default settings. A region was considered to have a proximal association with a gene if it was within 5kb upstream or 1kb downstream of the transcription start site (TSS). Regions outside this distance and up to 1000 kb from the TSS to the next gene proximal region were considered to have a distal association.

5.3. Results

Analysis of PR and ER α ChIP-seq data from T-47D breast cancer cells revealed 22,152 PR and 18,560 ER α binding regions with FDR < 5%. HOMER motif analysis on the top ranked 1,000 regions by peak score revealed the strong presence of a PRE motif (59.40%) and ERE motif (48.80%) (Table 5.1, 5.2). These were the most statistically significant motifs identified, in agreement with other studies (Lin et al., 2007; Kim et al., 2013a). In addition, in PR binding regions we found motifs for the transcriptional partners FOXA1 and AP-2 (TFAP2C) as other top ranked motifs. The transcription factor activator protein 2C (TFAP2C) was known to be involved in normal mammary development, differentiation, and oncogenesis (Cyr et al., 2015; Lal et al., 2013; Woodfield et al., 2010). Interestingly PR motifs were present in 344 (34.4%) of the 1,000 top ranked ER α binding regions. Consensus FOXA1 motifs were also detected in 27% of PR binding regions and 24% of regions bound by ER α . FOXA1, a member of the forkhead family of transcription factors, was known to bind and reconfigure condensed chromatin to enable the binding of other transcription factors (Bernardo and Keri, 2012). The presence of high quality (p-value < 1.00e-05) peaks and known conserved PR and ER α recognition sequences confirmed the success of the alignment and peak-calling process.

The size distribution of ER α (18,560 regions) and PR (22,152 regions) binding regions were visualised by drawing a histogram and box plot (Figure 5.1, 5.2). Mean PR binding region

size was 1508 bp with a median of 1336 bp. In contrast, ER α binding regions were on average half the size of PR binding regions, with a mean size of 601 bp and median 529 bp. Most PR binding regions (~94%) were greater than 1 kb, whereas most ER α binding regions (~95%) were less than 1kb. The longer PR regions may be due to longer input DNA fragment lengths in the original samples (Kharchenko et al., 2008; Landt et al., 2012) .




Motif	Name / Cell line	P-value	% of Targets Sequences with Motif
	PR(NR)/T47D	1e-123	59.40%
	FOXA1(Forkhead)/ LNCAP-FOXA1	1e-28	27.10%
	AP-2gamma(AP2)/ MCF7-TFAP2C	1e-10	13.70%

Table 5-1: Motif analysis of PR regions. Known motif analysis of PR top 1000 regions using Homer software.




Motif	Name / Cell line	P-value	% of Targets Sequences with Motif
	ERE(NR/IR3)/ MCF7-ERa	1e-474	48.80%
	FOXA1(Forkhead) / LNCAP-FOXA1	1e-22	24.30%
	PR(NR)/T47D-PR	1e-20	34.40%

Table 5-2: Motif analysis of ER α regions. Known motif analysis of ESR1 top 1000 regions.

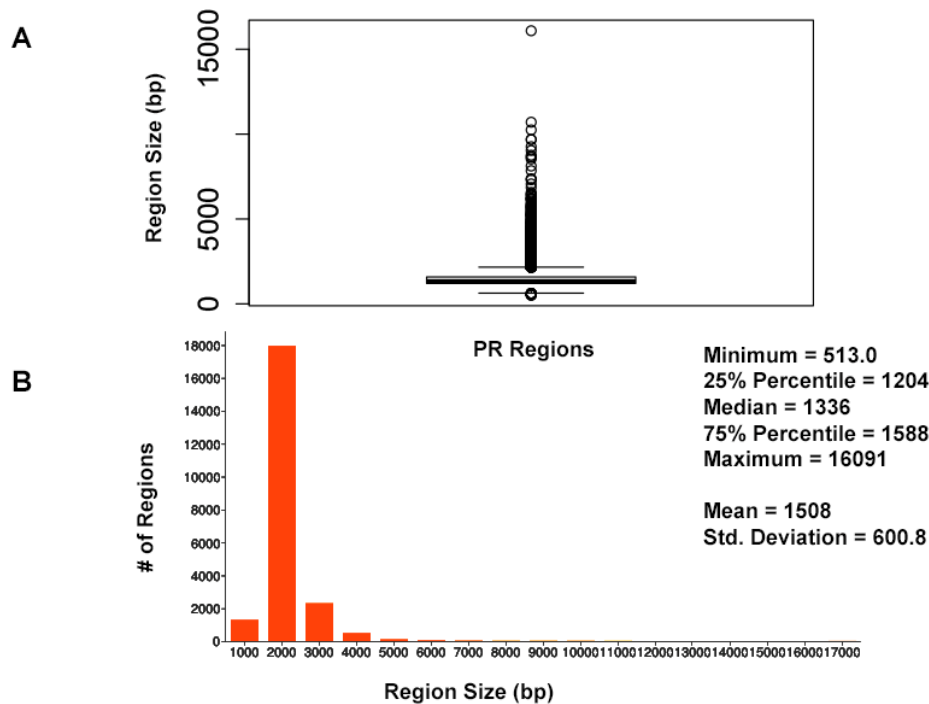


Figure 5-1: Distribution of PR binding region sizes. A) Box plot with mean and median information. B) Histogram of region sizes with bin size 1000 bp.

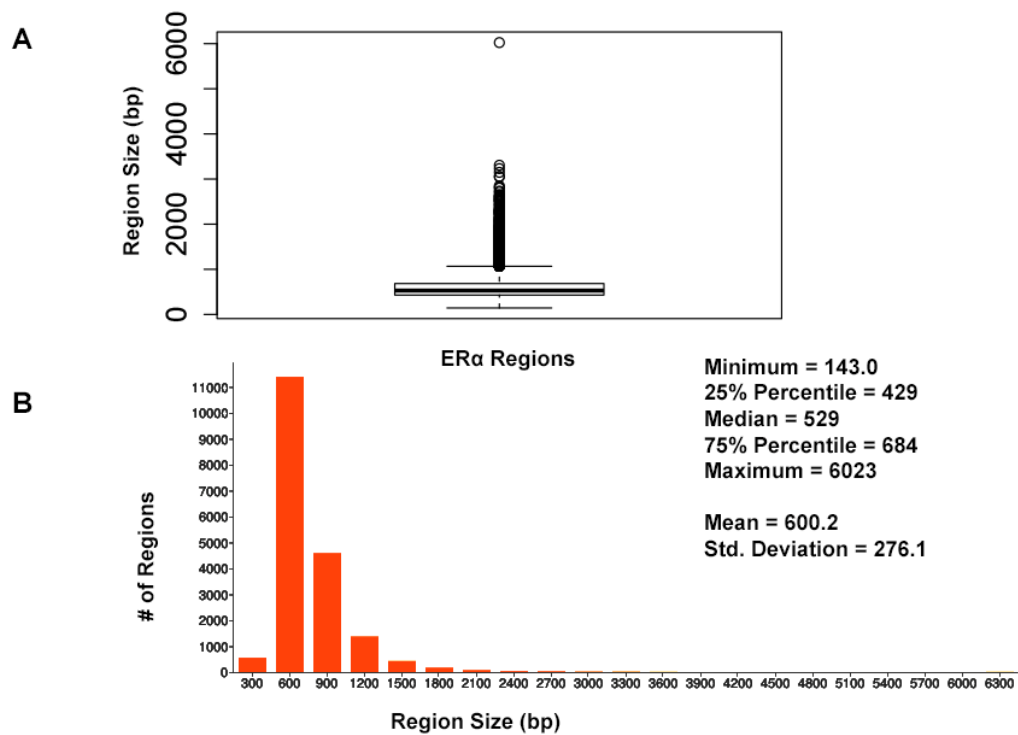


Figure 5-2: Distribution of ERα binding region sizes. A) Box plot with mean and median information. B) Histogram of ERα region sizes with bin size 200 bp.

5.3.1. Limited Overlap of ER α and PR Regions

Using BiSA, we identified that almost one quarter (23.6%) of ER α binding regions (4,344) overlap with 3,870 unique PR binding regions. This revealed that some long PR binding regions spanned more than one ER α binding region and the reverse was also true for large ER α binding regions. In total, we found 4,358 regions that were common to the two datasets. The Venn diagram in Figure 5.3-A shows this overlap between the two ligand-activated transcription factors. The 4,358 overlapping sections of the regions common to the two datasets were extracted and plotted for their region lengths (Figure 5.3-B). Out of 4,358 overlapping sections 4,279 (98.2%) were more than 100 bases long, suggesting a strong binding overlap between the two transcription factor data sets. An example of a shared ER α and PR binding region is shown in Figure 5.4. The 631 bp ER α binding region (red dotted lines) is completely contained within the 813 bp PR binding region (blue dotted lines) and the two regions share the peak centre location (Figure 5.4).

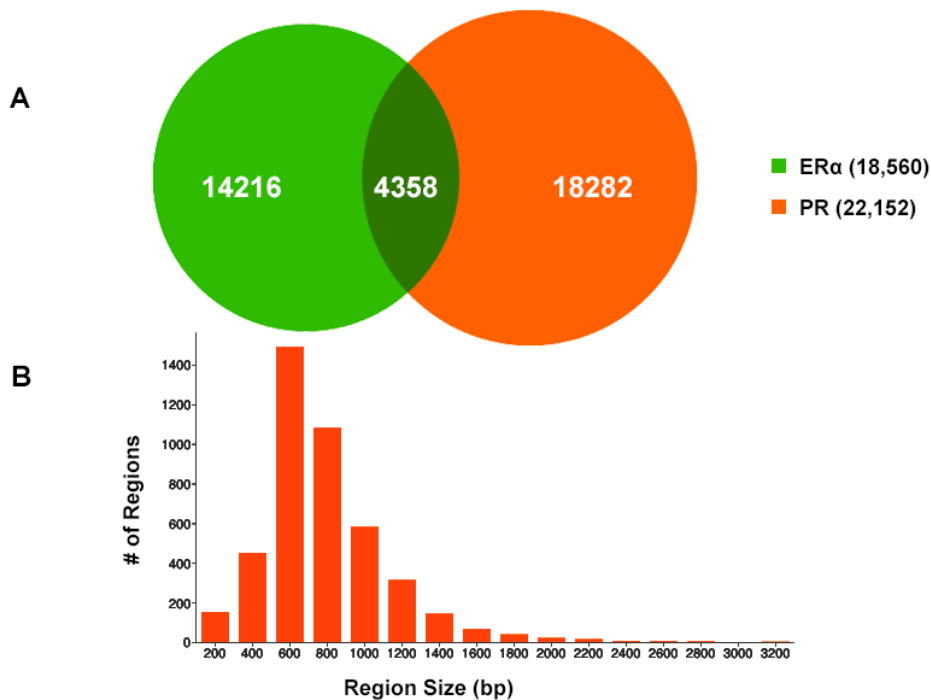


Figure 5-3: Visualisation of ER α and PR overlapping common regions.

A) Venn diagram showing overlap between ER α and PR data. The 4344 ER α binding regions overlap with 3870 unique PR binding regions making up 4358 overlapping sections. B) Region Sizes of 4358 overlapping sections of the common regions in ER α and PR datasets.

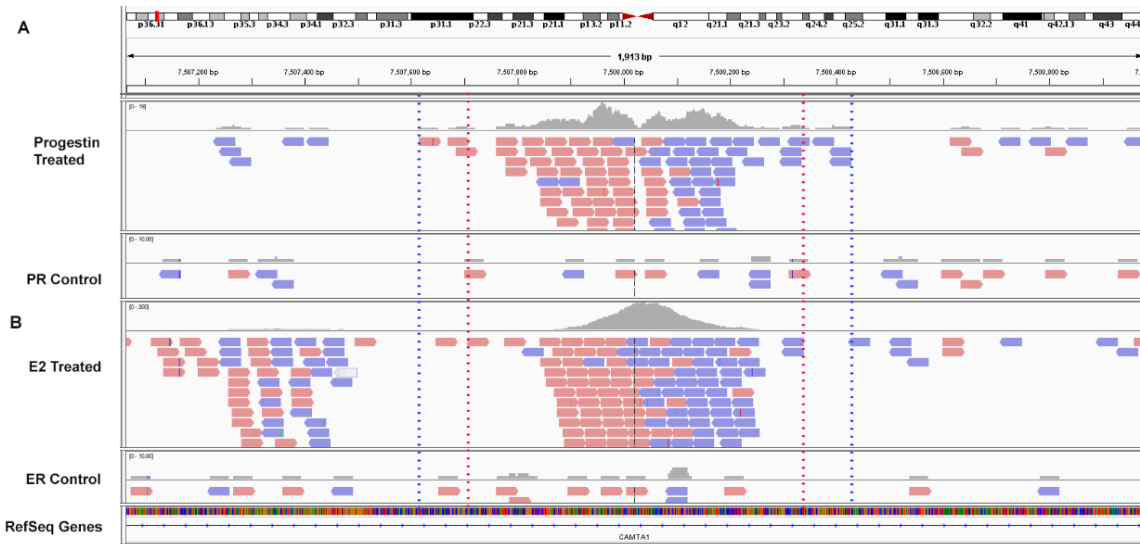


Figure 5-4: Example overlapping ER α -PR region. IGV (Integrative Genomics Viewer) snapshot of PR binding region at chr1:7507615-7508428 marked by blue dotted lines, ER α region is marked by red dotted lines. A) Progestin treated and control samples. The control as progestin treated input DNA sample. B) E2 treated and control sample. The red boxes are reads that mapped to forward strand and blue boxes are reads that mapped to reverse strand.

5.3.2. Statistical Analysis of ER α -PR Overlap

To determine whether the overlap between ER α and PR binding was statistically significant, statistical analysis was performed in BiSA, BITS and Genometricorr. In BiSA, using a whole genome domain and selecting the ER α cistrome as query and PR as reference revealed an overlap correlation value of 0.33. The value decreased to 0.26 when PR was selected as query and ER α as reference. This showed that, although a considerable proportion of ER α binding

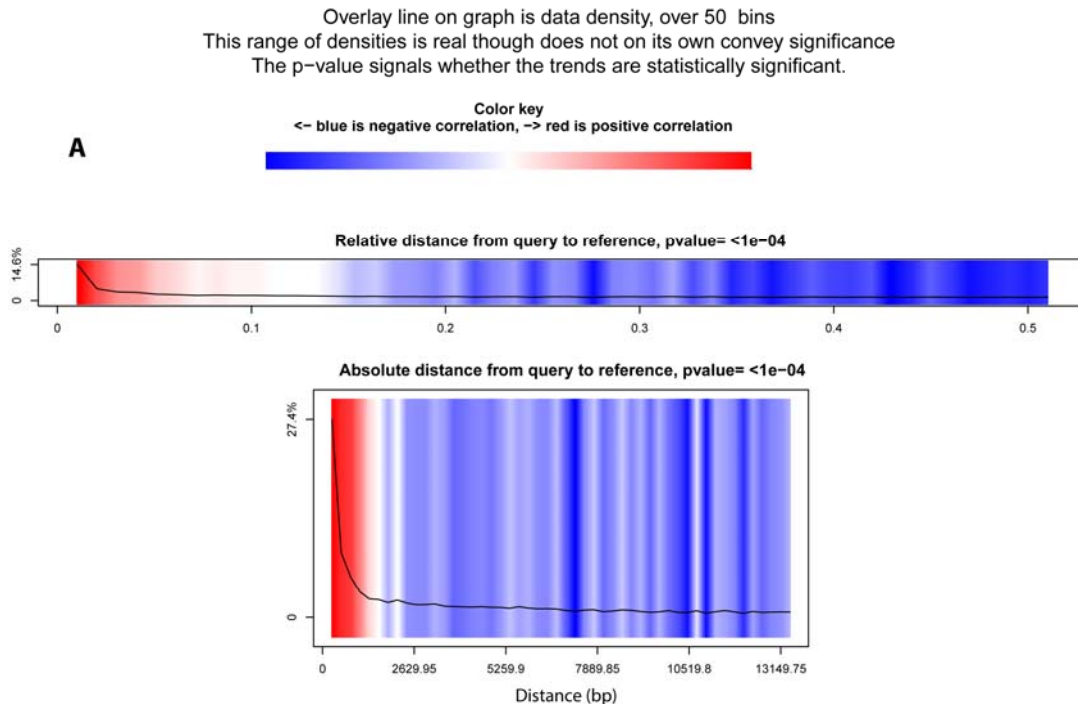
regions are also bound by PR, the two receptors do not cooperate for binding at all sites. To determine whether the significance of ER α -PR binding overlap was greater in functionally relevant genomic regions, we compared the level of binding overlap over a range of genomic domains from promoter proximal (within 500 b of a TSS) to more distal regions (Table 5.3). We found a low though consistent overlap correlation value (~ 0.3) whether promoter proximal or distal sites were included in the analysis (Table 5.3). To confirm that the OCV result is independent of the mean region sizes of the two datasets, we fixed the PR region sizes to 300 bases from each side of peak summits to match mean ER α region length (mean=601) and performed the OCV test again. This did not change the OCV (0.33) for the whole genome dataset, and there was negligible change in OCV observed for other domains (Table 5.3).

Using BITS and Genometricorr, we further investigated whether the spatial proximity correlation between PR and ER α binding was more significant than expected by chance. BITS Monte Carlo simulation reported that the spatial correlation of ER α and PR was statistically significant, with a p-value of 0.0001. Similarly Genometricorr's Local Correlation test, Absolute Distance test, Jaccard test and Projection tests also reported the spatial correlation between the two factors as statistically significant (p-value $\leq 1e-04$) (Figure 5.5). We repeated the tests for the 600bp fixed-width PR dataset and found no change in reported p-values from BITS or Genometricorr. This confirmed that a change in average region size between the two datasets does not affect the statistical analysis and demonstrated that the tendency for binding events for the two factors to be close to each other is statistically significant. Hence, the degree of overlap between the two factors was not significant, however, the spatial correlation of binding pattern was highly significant.

Domain	Overlap Correlation Value (OCV)			# of overlaps** / total ER α regions in domain
	Query = ER α Reference = PR	Query = PR Reference = ER α	Query = ER α Reference = PR (600 bp long)*	
Whole Genome	0.33	0.26	0.33	4344 / 18560
500 bp upstream, downstream of TSS	0.3	0.17	0.22	112 / 419
1 kb upstream, downstream of TSS	0.28	0.18	0.25	157 / 647
5 kb upstream of TSS	0.3	0.21	0.28	304 / 1224
5 kb upstream, downstream of TSS	0.31	0.22	0.3	522 / 2147
10 kb upstream, downstream of TSS	0.31	0.22	0.3	929 / 3666
5 kb upstream, downstream from 50kb upstream of TSS	0.29	0.21	0.28	449 / 1929
5 kb upstream, downstream from 100 kb upstream of TSS	0.31	0.24	0.3	514 / 2017
10 kb upstream, downstream from 100 kb upstream of TSS	0.31	0.23	0.3	878 / 3495

*Table 5-3: BiSA Overlap Correlation Value (OCV) testing. BiSA Statistical analysis of overlapping of ER α against PR dataset using different domain datasets. *PR binding regions are fixed to 600 bp long by cutting off 300 bp on both sides of peak summits. ** Number of overlaps in this column are reported by selecting ER α as query and PR as reference dataset.*

Therefore we conclude that, although there are a number of statistically significant shared binding sites in the ER α and PR datasets, and that ER α and PR often bind in proximity to each other, the observed overlap of the two factors is not strong enough for them to be considered as co-factors that consistently co-operate on shared binding regions. However, the close proximity of the binding regions for the two factors shows a spatial convergence and is statistically significant.



B

Results: All chromosomes

Overlap summary (Jaccard and projection tests)

Jaccard p-value: <1e-04

Query and reference intervals overlap significantly more than expected by chance, by Jaccard

Query midpoints and reference intervals overlap significantly more than expected by chance, by projection

Figure 5-5: Statistical significance test using Genometricorr. Genometricorr statistical significance analysis of ER α (query)-PR (reference). A) Relative and Absolute Distance Correlation tests are shown graphically. Overlay line is the data density when in blue section shows negative correlation while high density in red section shows positive correlation. B) Results from Jaccard and Projection tests are shown in text.

5.3.3. Motif Analysis

The 4,358 common sections of ER α -PR were searched for known motifs. Known motif analysis in these common sections revealed a strong presence of ERE, forkhead protein and PRE motifs. In Table 5.4, we list the top ranked motifs, ordered by p-value. A PRE motif was found in 41.88% (1,825) of the total 4,358 regions, which was much higher than the number of ERE motifs detected (14.3% (623) of the sequences). However, this may reflect the higher

stringency of the position specific scoring matrix used to identify ERE motif occurrence than the matrix used to find PRE motifs since the p-value for ERE motif detection ($1e-291$) was much stronger than the p-value for PRE motif occurrence in the dataset ($1e-179$). Secondly, the ERE motif came from an MCF7 dataset. The presence of FOXA1 motifs in these regions suggests that the factor facilitates the binding of ER α and PR on these regions as previously reported (Bernardo and Keri, 2012; Augello et al., 2011; Nakshatri and Badve, 2009). In addition AP-2 and TEAD4 (TEA) motifs were also identified in these regions and in the 1,000 top scoring PR binding regions. AP-2 has a known role in normal mammary development and breast cancer (Cyr et al., 2015; Lal et al., 2013; Woodfield et al., 2010). TEAD4 has also been shown to be co-expressed with other oncogenes and is correlated with poor prognosis (Xia et al., 2014; Mesrouze et al., 2014; Lim et al., 2014). The presence of the related motifs in the ER α -PR shared regions as well as in regions that bind uniquely ER α or PR suggests that AP-2 and/or TEAD play a key role for both receptors and could be important in facilitating cooperation between the two nuclear receptors.

Using Homer, we also looked at relative position distributions of these motifs (Figure 5.6). We found that the motifs converge around the centres of the peaks, supporting their biological significance as primary binding events.

We also performed *de novo* motif analysis that identified a dominant ERE element in the common section a canonical forkhead target sequence RYAAAYA. The symbol R in IUPAC (International Union of Pure and Applied Chemistry) codes represents the occurrence of either A or G and Y represents either C or T.






Motif	Name / Cell line	P-value	% of Targets Sequences with Motif
	ERE(NR/IR3)/ MCF7-ERa	1e-291	14.30%
	FOXA1(Forkhead)/ LNCAP-FOXA1	1e-249	35.11%
	PR(NR)/ T47D-PR	1e-179	41.88%
	AP-2gamma(AP2)/ MCF7-TFAP2C	1e-122	20.38%
	TEAD4(TEA)/ Tropoblast-Tead4	1e-86	17.97%

Table 5-4: Known motif analysis of ERa and PR overlapping common regions. Top ranked known motif analysis of ERa-PR common sections (4358 regions)

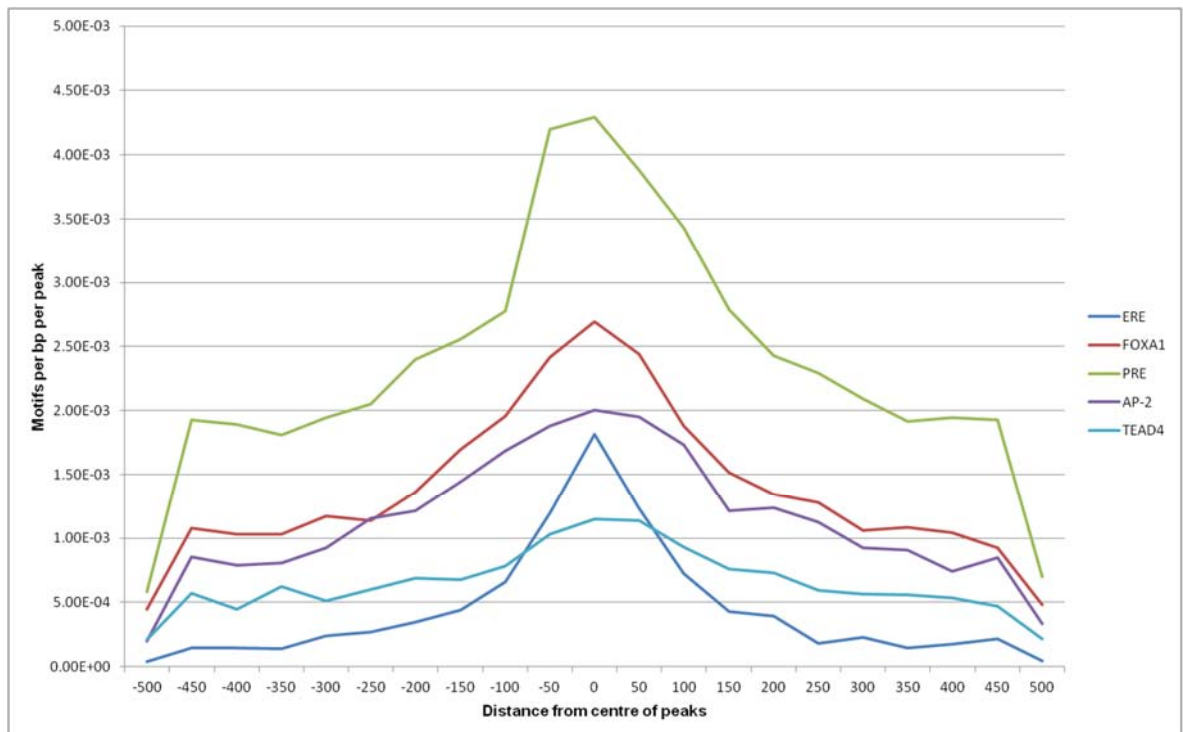


Figure 5-6: Motif position distributions in ERa-PR overlapping regions. Frequency distribution of ERE, FOXA1, PRE, AP-2 and TEAD4 motifs around centre of peaks using 50

bp bin size.

The presence of forkhead target sequence in known and *de novo* motifs confirms the fact that the factor facilitates the binding of ER α and PR on these regions. *De novo* motif analysis also revealed AP-2/TFAP2C, TEAD4, GRHL2(CP2), and Ets1-distal (ETS), however, interestingly *de novo* analysis didn't pick up the PRE (Table 5.5). Therefore we postulate that PR binding on some of these regions could be due to protein-protein tethering. AP2 and TEAD4 again came up as significant *de novo* motifs supporting their key role in these common regions.

Rank	Motif	P-value	% of Targets	Best Match/Details
1		1e-370	16.27%	MA0112.2_ESR1
2		1e-280	39.19%	MA0031.1_FOXD1
3		1e-138	37.70%	AP-2gamma(AP2)/MCF7-TFAP2C
4		1e-128	17.05%	TEAD4(TEA)/Tropoblast-Tead4
5		1e-101	14.20%	GRHL2(CP2)/HBE-GRHL2
6		1e-87	33.30%	Ets1-distal(ETS)/CD4+-PolII

Table 5-5: De novo motif analysis of ER α and PR overlapping common regions. Top ranked de novo motif analysis of ER α -PR common sections (4358 regions)

We extracted the exact locations of ERE and PRE motifs from 4,358 ER α -PR common sections. We identified 8,259 PR motif locations and 1,831 ER α motif locations, ~4.5 times more PR motifs than ER α motifs. However, lower P-value of PR motifs (Table 5.4) could be

due to higher sequence variability than ER α motif sequences. We also identified 598 motif locations sharing at least 1 bp in common (Figure 5.7). Of these 598 locations 285 locations were exactly the same. The sequences of these 285 locations were extracted using the UCSC Genome Browser and a sequence logo was drawn using Web Logo (Crooks et al., 2004) to visualise the pattern of these sequences (Figure 5.8). The difference in the middle part of this logo against PRE or ERE was quite noticeable, for example at the 7th position A was dominant with almost equal probability of T, G or C nucleotides; while in case of ERE logo it was almost always A and in PRE it was dominant A with equally shared probability between T or G.

5.3.4. ER α -PR Common Regions Interact on Enhancer Regions

We further investigated whether ER α and PR interact on enhancer (H3K4me1) or on promoter (H3K4me3) marks. Enrichment for monomethylation of histone H3 lysine 4 (H3K4me1) is one of well studied chromatin signatures; these regions are known to involve in increasing transcription process. Whereas, enrichment for trimethylation of histone H3 lysine 4 is tightly associated with promoters of active genes (Smith and Shilatifard, 2014). We observed that most ER α -PR overlapping sections (3018, 69.3%) overlap with H3K4me1 marks (Figure 5.9), while only 201 (4.6%) regions overlapped with H3K4me3 marks. BiSA statistical analysis revealed a moderately significant overlap correlation value (OCV) of 0.51 when ER α -PR common regions (4,358) were selected as query and H3K4me1 (73,263 regions) selected as reference dataset against a whole genome domain. Therefore this confirmed that ER α and PR binding on these regions were facilitated by enhancer marks (Gadaleta and Magnani, 2014).

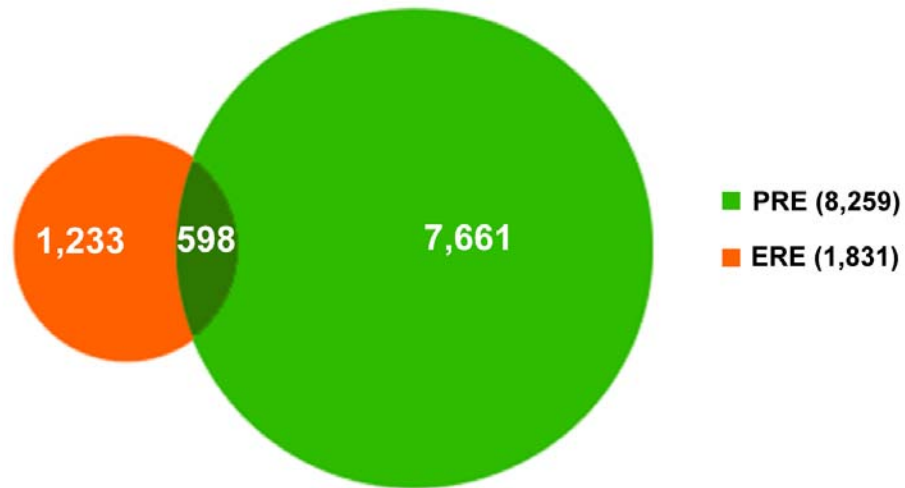


Figure 5-7: Overlapping of 1,831 ERE and 8,259 PRE motif locations in 4358 common ERa-PR regions.

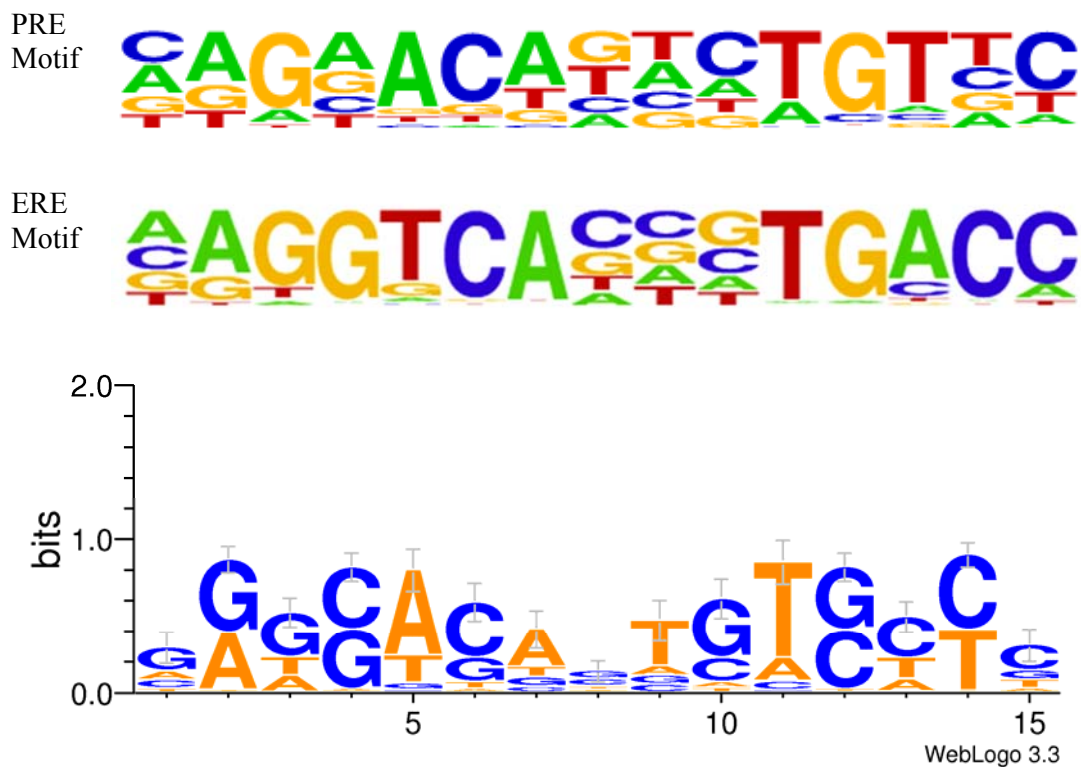


Figure 5-8: Comparison of PRE and ERE motifs with sequence logo generated from sequences of 285 common ERE-PRE motif locations. Error bars represent confidence interval in variation of letter height.

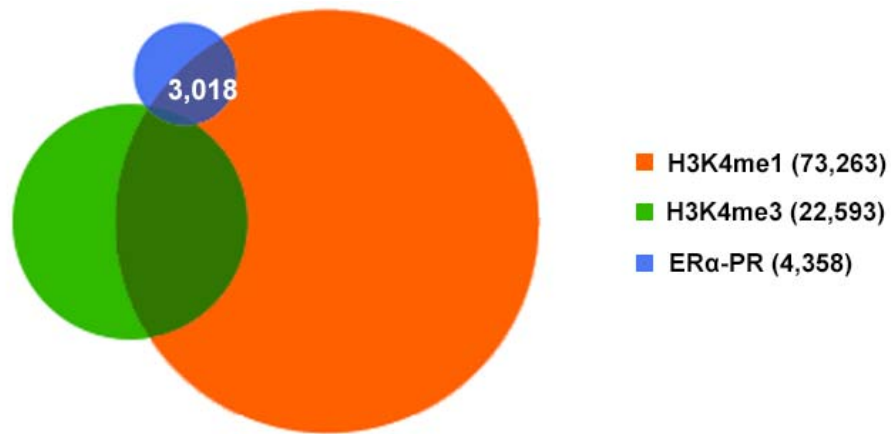


Figure 5-9: Venn diagram showing an overlap of H3K4me1, H3K4me3 and ERα-PR common regions.

5.3.5. Bioinformatics Enrichment Analysis of ERα-PR Common Regions

We used GREAT (Genomic Regions Enrichment of Annotations Tool) (McLean et al., 2010) to interpret the functional role of 4,358 ERα-PR common regions. Using GREAT default parameter as described in Methods, regions were annotated. GREAT revealed that only 34 regions (~0.8%) are not associated with any gene and 3,687 (~85%) regions are associated with 2 genes (Figure 5.10). Most of the regions were found to be distal binding events while 405 (~9%) regions are within 5kb of transcription start sites (TSS). Region to gene association revealed MYC has the maximum number of regions linked to this gene (26 regions). The known role of the estrogen-induced MYC oncogene in breast cancer (Wang et al., 2011a; Orr et al., 2012) confirms a biologically relevant regions-to-gene association. PGR was also among the top 10 genes identified with the largest number of associated regions (Table 5.6). Gene ontology enrichment analysis of the genes associated with common regions revealed epithelial cell development as the most significant biological process (Table 5.7). Epithelial cell development was linked to 30 genes associated with 120 regions out of which 4 regions were within 5kb of a TSS. Pathway Commons, a meta-database of public biological pathway information (Cerami et al., 2006), revealed the ERα signalling network as the most

significant term (p-value=5.7e-37) where 137 regions were found regulating 24 genes associated with this pathway (Table 5.8). The FOXA1 transcription factor network and IL6-dedicated signalling events were also significant terms (p-value 1.6e-19 and 2.6e-17). Mouse phenotype analysis revealed two breast cancer related ontologies (abnormal mammary gland epithelium physiology and abnormal mammary gland development) as the most significant terms. There were 32 regions associated with 5 genes linked to abnormal mammary gland epithelium physiology and 189 regions associated with 52 genes linked to mammary gland development.

Genes	Total Regions
MYC	26
KCNMA1	23
TRPS1	22
TYRP1	18
EIF3H	18
MED13L	18
MPDZ	18
DSCAM	17
KIAA0182	16
PGR	15

Table 5-6: Genes associated with ER α -PR common regions.

GO ID	Description	Binomial Rank	Binomial P-Value	Binomial Bonferroni P-Value	Hypergeometric P-Value	Hypergeometric Bonferroni P-Value	Observed Genes	# of Regions	Genes
0002064	Epithelial cell development	32	5.29E-21	4.63E-17	7.35E-06	6.44E-02	30	120	AGT,B4GALT1,BASP1,BCL11B,BMP4,CITED1,COL18A1,DICER1,ESR1,FOXA1,GATA3,GDNF,GJA1,GREM1,GSTM3,HEG1,KCNE1,ONECUT1,PDE4D,PGR,RAC1,SHROOM3,SPDEF,TFAP2A,TFAP2C,TP63,VEZF1,WNT7B,WT1,XBP1
0051897	Positive regulation of protein kinase B signaling cascade	38	2.99E-20	2.62E-16	2.16E-03	1.00E+00	18	83	ANGPT1,CCR7,EGFR,F3,GATA3,IGF1,IGF1R,IGFBP5,IL6,INSR,MTDH,NOX4,PTPRJ,SPRY2,TCF7L2,TGFBR1,THBS1,TSPYL5
0045834	Positive regulation of lipid metabolic process	134	3.25E-13	2.84E-09	4.11E-04	1.00E+00	30	92	ABCG1,AGT,APOA1,CCR7,CYP17A1,EPHA8,FGF1,FGF2,FGFR3,FLT1,GHSR,IGF1R,IRS1,IRS2,KIT,LDLRAP1,MID1IP1,NOD2,PDGFB,PNPLA2,PPARA,PPARGC1A,PRKAA1,PRKCD,PRKCE,RAC1,SOBBS1,SREBF1,VAV2,VAV3
043551	Regulation of phosphatidylinositol 3-kinase activity	140	5.98E-13	5.24E-09	1.10E-04	9.64E-01	13	43	CCR7,EPHA8,FGF2,FGFR3,FLT1,IRS1,KIT,NOD2,PDGFB,PIK3R1,RAC1,VAV2,VAV3
0043552	Positive regulation of phosphatidylinositol 3-kinase activity	165	4.42E-12	3.87E-08	1.67E-04	1.00E+00	12	35	CCR7,EPHA8,FGF2,FGFR3,FLT1,IRS1,KIT,NOD2,PDGFB,RAC1,VAV2,VAV3
0090218	Positive	171	7.76E	6.80E-	3.15E-04	1.00E+00	12	35	CCR7,EPHA8,FGF2,FGFR3,FLT1,IRS1,KIT,NOD2,

	regulation of lipid kinase activity		-12	08		0			PDGFB,RAC1,VAV2,VAV3
0050731	Positive regulation of peptidyl-tyrosine phosphorylation	195	2.92E-11	2.56E-07	1.97E-04	1.00E+00	34	98	ADNP,AGT,ANGPT1,CD44,EFNA5,EHD4,FGF10,FGFR3,GHR,HES1,HGF,IGF1,IL12A,IL12B,IL15,IL20,IL6,IL6ST,ITGB1,JAK2,KIT,KITLG,LIF,LOC284889,NOD2,NRP1,OSM,PAK2,PDGFB,PTK2B,RICTOR,SYK,TKN2,VEGFA
0043550	Regulation of lipid kinase activity	211	9.74E-11	8.53E-07	1.53E-03	1.00E+00	13	43	CCR7,EPA8,FGF2,FGFR3,FLT1,IRS1,KIT,NOD2,PDGFB,PIK3R1,RAC1,VAV2,VAV3
0060740	Prostate gland epithelium morphogenesis	261	1.84E-09	1.61E-05	6.07E-04	1.00E+00	13	61	AR,BMP4,CD44,ESR1,FGFR2,FOXA1,FRS2,GLI2,IGF1,IGF1R,NOG,SOX9,TP63
0060512	Prostate gland morphogenesis	273	2.39E-09	2.10E-05	9.81E-04	1.00E+00	13	61	AR,BMP4,CD44,ESR1,FGFR2,FOXA1,FRS2,GLI2,IGF1,IGF1R,NOG,SOX9,TP63

Table 5-7: Top 10 Gene Ontology (GO) biological process associated with ER α -PR common regions. P-values are calculated by binomial and hypergeometric tests and corrected by Bonferroni Correction. Binomial Rank is based on the Binomial P-value, lowest P-value gets higher rank.

PC ID	Description	Binomial Rank	Binomial P-Value	Binomial Bonferroni P-Value	Hypergeometric P-Value	Hypergeometric Bonferroni P-Value	Observed Genes	# of Regions	Genes
517105	Validated nuclear estrogen receptor alpha	40	5.78E-37	9.28E-34	1.83E-03	1.00E+00	24	137	AP1B1,ATP5J,C10orf12,CALCOCO1,CCND1,CD82,CEBPB,COL18A1,CTSD,DSCAM,EBAG9,ESR1,GREB1,HDAC4,HSF2,LMO4,MTA1,MYC,NCOA3

	network								,NCOR2,NRIP1,PGR,PRDM15,XBP1
517173	FOXA1 transcription factor network	46	1.60E-19	2.58E-16	1.06E-06	1.70E-03	24	96	AP1B1,AR,ATP5J,CEBPB,COL18A1,DSCAM,ESR1,FOS,FOXA1,FOXA2,FOXA3,NCOA3,NFIA,NFIB,NFIC,NR2F2,NRIP1,PISD,PRDM15,SCGB1A1,SERPINA1,SFTPA1,SFTPD,XBP1
517023	IL6-mediated signaling events	51	2.67E-17	4.29E-14	2.27E-04	3.65E-01	21	83	BCL2L1,CEBPB,CEBPD,FOS,FOXO1,GAB2,HCK,IL6,IL6ST,IRF1,JAK1,JAK2,LMO4,MAPK14,MITF,MYC,PIAS1,PIAS3,PIK3R1,PRKCD,RAC1
517048	FOXA transcription factor networks	53	4.20E-17	6.75E-14	8.24E-07	1.32E-03	37	114	ABCC8,ALAS1,AP1B1,APOA1,AR,ATP5J,CEBPA,CEBPB,CEBPD,COL18A1,CREB1,DSCAM,ESR1,FOS,FOXA1,FOXA2,FOXA3,FOXF1,HADH,KCNJ11,NCOA3,NF1,NFIA,NFIB,NFIC,NR2F2,NR3C1,NRIP1,PISD,PRDM15,SCGB1A1,SERPINA1,SFTPA1,SFTPD,TAT,TFRC,XBP1
517138	C-MYB transcription factor network	55	2.42E-16	3.89E-13	2.50E-04	4.02E-01	32	122	ATP2B1,BIRC3,CBX4,CCND1,CD34,CDK6,CEBPA,CEBPB,CEBPD,COL1A2,GATA3,HES1,HIPK2,IQGAP1,KIT,KITLG,LEF1,MAD1L1,MAT2A,MYB,MYC,NLK,PIAS3,PPP3CA,RAG2,SND1,TAB2,TFEC,TOM1,UBE2I,YEATS4,ZFH3
485310	Effects of PIP2 hydrolysis	62	3.31E-14	5.32E-11	1.67E-04	2.68E-01	12	51	DAGLA,DAGLB,DGKB,DGKH,DGKI,DGKK,PRKCD,PRKCE,PRKCH,PRKCQ,TRPC6,TRPC7
485288	Platelet activation, signaling and aggregation	64	7.00E-13	1.13E-09	9.66E-04	1.00E+00	45	128	ADRA2A,AP2B1,AP2S1,APOA1,ARRB1,BCAR1,CALM1,CALM2,CAP1,DAGLA,DAGLB,DGKB,DGKH,DGKI,DGKK,FYN,GAS6,GNA13,GNA14,GNAAQ,GNB1,GRIP2,IGFBP3,LAMP2,LAT,MAPK14,MAPK3,P2RY1,PDPK1,PIK3R1,PRKCD,PRKCE,PRKCH,PRKCQ,PTK2,PTPN1,RAPGEF3,RHOB,SEPT5,SYK,TRPC6,TRPC7,VAV2,VAV3,YWHAZ
517065	HIF-1-alpha transcription factor network	65	1.55E-12	2.49E-09	1.92E-04	3.08E-01	27	82	ABCG2,ADM,ALDOA,BHLHE40,BNIP3,CITED2,CP,CREB1,CXCL12,EDN1,EGLN1,EGLN3,FOS,HIF1A,HK1,HMOX1,ID2,ITGB2,NDRG1,NOS2,NT5E,PFKFB3,RORA,SLC2A1,TFF3,TFRC,VEGFA

517031	Integrins in angiogenesis	74	1.58E-10	2.54E-07	7.45E-04	1.00E+00	25	75	BCAR1,CD44,CSF1,FGF2,FN1,FOS,IGF1,IGF1R,IRS1,KDR,MAP3K1,MAPK3,MAPK8,NFKBIA,PIK3R1,PTK2,PTK2B,PXN,RAC1,ROCK1,SDC1,SYK,TGFBR2,VAV3,VEGFA
--------	---------------------------	----	----------	----------	----------	----------	----	----	----------------------------------------------------------------------------------------------------------------------------------------

Table 5-8: Top 10 Pathway Commons terms associated with ERα-PR common regions.

MP ID	Description	Binomial Rank	Binomial P-Value	Binomial Bonferroni P-Value	Hypergeometric P-Value	Hypergeometric Bonferroni P-Value	Observed Genes	# of Regions	Genes
0010172	Abnormal mammary gland epithelium physiology	19	1.20E-22	8.78E-19	2.02E-03	1.00E+00	5	32	CTNNA1,DDR1,KDM4B,NCOA3,PGR
0000628	Abnormal mammary gland development	44	1.86E-19	1.36E-15	1.09E-06	7.94E-03	52	189	AHR,AR,AREG,ARHGAP5,B4GALT1,BCL2L11,CCND1,CD44,CDH1,CEBPB,CITED1,CSF1,DDR1,ELF5,ESR1,FGF10,FGFR2,FKBP4,FOXA1,GATA3,GJA1,GLI2,GLI3,HPRT1,IGF1,IL6,IL6ST,ITGB1,JAK2,KDM4B,LEF1,LIF,MKL1,MSX1,MSX2,MYBL1,NCOA3,NOS2,NR3C1,NRG1,NRG3,NTN1,PGR,PHB,PLGLB1,PRLR,TGFA,TP63,UBE3A,XDH,ZFPM2
0002098	Abnormal vibrissa morphology	46	3.85E-19	2.81E-15	1.36E-06	9.94E-03	34	126	ACD,AREG,ATP7B,BARX2,BMP7,CTNNA1,DICER1,DLX6,EGFR,FGFR2,FOXC1,GATA3,GLI3,HOXC13,INHBA,INHBB,KITLG,KRT17,LAMP2,LBR,LEF1,MECP2,MOCS1,MSX2,POU3F4,RIPK4,SGK3,SPINK5,ST14,TCF

									7L2, TGFA, TP63, TRPS1, VDR
0002842	Increased systemic arterial blood pressure	50	1.61E-18	1.17E-14	2.86E-04	1.00E+00	40	140	ADM, ADORA2A, AGT, CD44, CD47, CHGA, CORIN, CTH, CYP4A11, DDAH1, DLL1, DRD5, EDN1, EGFR, GDNF, GPER, HPRT1, HSD11B2, IER3, IGF1, IRS1, IRS2, KCNJ11, KCNK9, KCNMA1, NEDD4L, NOS2, NR3C1, PODXL, PPM1L, PRKG1, PTGIS, SCNN1B, SPECC1L, THBS1, TRPC6, VAV2, VAV3, VDR, WNK1
0006382	Abnormal lung epithelium morphology	68	6.95E-17	5.07E-13	1.31E-05	9.56E-02	44	155	ABCA12, ARRB1, BID, BMPER, CEBPA, CELSR1, CITED2, CTGF, DICER1, EGFR, EPAS1, ERFFI1, EYA1, FOXA2, GLI2, GREM1, HES1, HIF1A, HPS1, IGF1, ITCH, KEAP1, KLF5, LIF, LMO7, MAPK14, NDST1, NEUROD1, NFIB, NR3C1, PGGT1B, RAB38, RUNX3, S1PR3, SCGB1A1, SFTPD, SOX2, TCF21, TGFBR1, THBS1, TMEM38B, TRPS1, VEGFA, WNT7B
0008372	Small malleus	97	3.67E-15	2.68E-11	1.62E-03	1.00E+00	6	37	GDF6, HOXA1, MSX1, MSX2, MYC, TBX1
0010900	Abnormal pulmonary interalveolar septum morphology	105	6.38E-15	4.65E-11	2.67E-03	1.00E+00	21	87	ABCA12, B4GALT1, CAV2, DUSP1, EGFR, EPAS1, ERFFI1, HCK, HIF1A, IGF1, KL, LIF, MAN1A2, MAPK8, NDST1, NR3C1, PKDCC, SFTPD, TMEM38B, TRPS1, VEGFA
0001284	Absent vibrissae	120	4.50E-14	3.28E-10	2.39E-03	1.00E+00	12	60	CTNNA1, HOXC13, INHBA, INHBB, KRT17, LAMP2, LEF1, RIPK4, ST14, TCF7L2, TP63, TRPS1
0001179	Thick pulmonary interalveolar septum	123	4.90E-14	3.57E-10	6.92E-04	1.00E+00	20	79	ABCA12, B4GALT1, CAV2, DUSP1, EGFR, EPAS1, ERFFI1, HCK, HIF1A, IGF1, LIF, MAN1A2, MAPK8, NDST1, NR3C1, PKDCC, SFTPD, TMEM38B, TRPS1, VEGFA
0001881	Abnormal mammary gland	127	6.35E-14	4.63E-10	3.02E-03	1.00E+00	30	98	AR, AREG, ATP7B, B4GALT1, CCND1, CDH1, CEBPB, CSF1, CTNNA1, DDR1, EGFR, ELF5, E

	physiology								SR1,FOXB1,GJA1,GUSB,HIF1A,ID2,INHBB, JAK2,KDM4B,LIF,MKL1,NCOA3,NOS2,PG R,PLGLB1,PRLR,TGFA,XDH
--	------------	--	--	--	--	--	--	--	-----------------------------------------------------------------------------------------------------

Table 5-9: Top 10 mouse phenotypes associated with ER α -PR common regions.

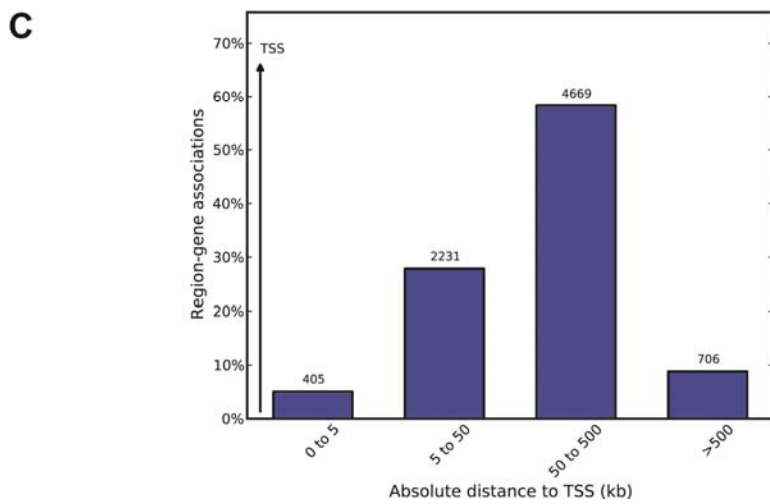
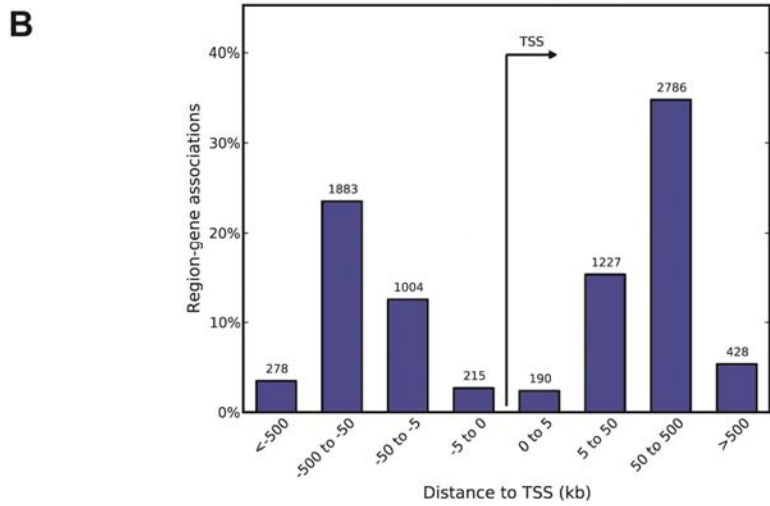
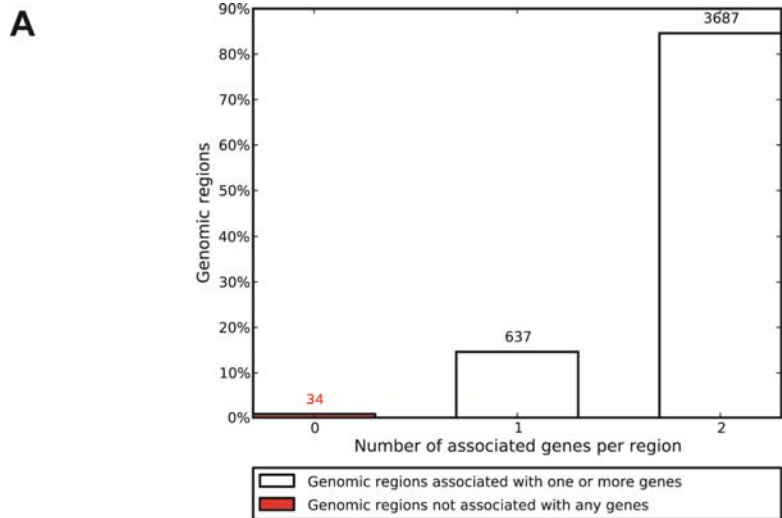


Figure 5-10: ER α -PR common regions-gene association. A) Number of associated genes per region. B) Region-gene association binned by orientation and distance to TSS. C) Region-gene association binned by absolute distance to TSS.

5.3.6. Differential Gene Expression Analysis by ER α /PR Binding

Clarke and Graham (Clarke and Graham, 2012) reported 1005 regulated transcripts by PR binding while Gertz et al. (Gertz et al., 2012) published 920 regulated transcripts by ER α binding in T47D cells. We found that 63 regulated genes were in common in the two gene expression studies. Of these 63 common regulated genes, 20 were up-regulated while 14 genes remained down-regulated in both treatments (Table 5.10).

Genes	Estrogen treatment	Progesterone treatment
ABCC12, ACOT1, ACOT2, AGR3, AZGP1, BNIP3, CITED4, CLIC6, CMTM7, GADD45B, HK2, IL6ST, KRT15, KRT16, RAB4B, RUNX1, SEC14L2, SERPINA3, SLC16A3, SPAG4	Up \uparrow	Up \uparrow
C3orf57, CLDN1, EPHA4, TGFB3	Down \downarrow	Down \downarrow
ACOX2, CADM1, CXCL12, FGD3, KCTD6, NPEPPS, OLFM1, PDZK1, PGR, P XK, RARA, RBBP8, STC2, TEX14	Up \uparrow	Down \downarrow
ARL4A, ARL4D, CDKN2B, CORO2A, CYFIP2, DLG2, DLX1, EFNA1, FAM107B, FBXO32, FZD4, ITPKA, KCNB1, MSX2, MTERFD3, MTSS1, NANOS1, RCAN1, SLC7A8, SOX9, TM4SF1, TNFRSF21, TRIM29, TSC22D3, VTCN1	Down \downarrow	Up \uparrow

Table 5-10: Common differentially expressed transcripts by the treatment of estrogen or progesterone.

5.3.7. Gene Expression Regulation due ER α -PR Common Regions

To investigate the regulation due to regions that were common in ER α and PR, we annotated the 4,358 ER α -PR overlapping regions for all genes with transcription start sites (TSS) within 50 kb of a shared region using BiSA and identified 3,338 genes nearby to the binding regions. Cross comparing these genes with respective ER α and PR gene expression data, we found that 264 genes were regulated by progestins and 218 genes were reported to be estrogen-regulated in T47D cells. 35 genes were present in the both ER α and PR expression datasets. We further investigated the functional relationships of these

differentially regulated transcripts using The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 (Huang da et al., 2009). DAVID functional annotation clustering for each set of transcripts that were common in gene expression data and BiSA reported genes with TSS within 50 kb of binding regions, revealed regulation of apoptosis or anti-apoptosis as one of the top functional annotation cluster (Table 5.11).

Group of Transcripts	Functional Annotation	Count	P-Value
Regulated by ERα-PR Common Regions			
264 genes regulated by progesterin and ER α -PR common regions	Negative regulation of cell differentiation	12	4.0E-4
	Negative regulation of myeloid cell differentiation	5	9.8E-4
218 genes regulated by estradiol and ER α -PR common regions	Identical protein binding	19	1.2 E-3
	Apoptosis	17	4.5 E-3
	Programmed cell death	17	5.2 E-3
35 common genes regulated by progesterin and estradiol	Monocarboxylic acid binding	3	6.9 E-3
	Apoptosis	5	4.4E-3
	Programmed cell death	6	1.5E-2

Table 5-11: Top DAVID functional annotation of estrogen and progesterin regulated transcripts that were associated with ER α -PR shared binding regions.

The identification of processes related to cell differentiation and apoptosis as top biological processes due to transcripts that were regulated by ER α -PR shared regions suggests that

the ER α -PR overlapping regions had an important role in regulating genes that involved in facilitating programmed cell death. Therefore we concluded that ER α -PR overlap is biologically relevant.

5.4. Discussion

The BiSA database provides a good starting point for studying overlapping binding of a range of transcription factors from a comprehensive collection of published studies (Khushi et al., 2014). The datasets available in BiSA represent the original genomic locations identified in the published studies from which they are sourced. Although the same standard pipeline has often been applied, it must be acknowledged that differences in read alignment algorithms (Lunter and Goodson, 2011; Kerpedjiev et al., 2014) and the use of a variety of peak-caller programmes (Ladunga, 2010; Pepke et al., 2009; Wilbanks and Facciotti, 2010) has an impact on downstream analysis, largely due to differences in stringency that affect the number of genomic regions identified. Our initial investigation of the overlap in ER α and PR binding in T-47D cells, utilizing the published binding regions, revealed an overlap of ~27% of ER α binding regions with the published PR cistrome (data not shown). This suggested an interesting functional relationship between the receptors, which justified further study. To perform a more rigorous exploration of their overlapping binding patterns, we reanalysed the raw ER α and PR ChIP-seq data using a standardized pipeline. This illustrates the value of BiSA as an easy to implement first pass tool to investigate potential functional relationships in transcription factor binding and epigenomic datasets.

The BiSA statistical overlap correlation value (OCV) represents a statistical summary value of the set of p-values calculated by the IntervalStats tool and reflects the overall correlation of two binding site datasets. IntervalStats calculates a p-value for each query region against the closest reference region within the given domain. It is designed to

identify factors that target the same genomic locations. As described in examples in our previous study (Khushi et al., 2014) the OCV should be greater than 0.5 for partner factors, reflecting a statistically significant correlation between two binding patterns. For example the OCV for known partners, FOXA3 (query) to FOXA1 (reference) was 0.72 (Motallebipour et al., 2009b). Similarly the OCV for CTCF (query) and SA1 (reference), which are known to co-locate on DNA, was 0.82 (Schmidt et al., 2010). Therefore the lower OCV for ER α -PR suggests that the majority of ER α and PR binding events are independent of each other, however, the OCV test does not challenge the biological co-occurrence of binding of the two factors on the reported regions where IntervalStats reports a statistically significant p-value. A consistent overlap was found both proximal and distal to gene promoters (Table 5.3). It is acknowledged that gene expression is regulated through interaction at a number of cis-regulatory elements, which includes promoters and enhancers. Moreover, enhancers can spread over a range of distances from the TSS. Therefore, the detection of binding sites over a range of distances and locations is to be expected (Calo and Wysocka, 2013; Bulger and Groudine, 2011a). This spatial correlation between the two factors is identified as statistically significant by Monte Carlo simulation using BITS, Relevant Distance, Absolute Distance, Jaccard and Projection tests using Genometricorr. Therefore, the regions from the two factors are found in close proximity more often than expected by chance although they do not exactly overlap. Therefore the consistent OCV observed using various domains and statistically significant spatial convergence suggest that the consistent overlap may have biological significance. Although not all sites overlapped, many of the shared ER α and PR binding regions were highly statistically significant binding sites for both receptors, as determined by a strong p-value and low FDR value in MACS, suggesting that these are biologically valid binding regions for these receptors and that their overlap reflects converging function on a subset of gene targets.

ER α and PR overlapping regions found to significantly colocalize with enhancer regions (H3K4me1). It would be nice to compare the results with H3K27 acetylation marks, however, at this stage no such data is available publicly.

In recent years a number of studies have published ER α binding regions in the MCF-7 cell line (Grober et al., 2011; Schmidt et al., 2010; Welboren et al., 2009; Tsai et al., 2010; Hurtado et al., 2008; Gu et al., 2010; Joseph et al., 2010; Hu et al., 2010). However only two studies have published ER α data in T47D cells (Joseph et al., 2010; Gertz et al., 2012). We chose to study the Gertz et al. 2012 dataset because using data from the Joseph et al. study we called only 1,817 peaks with FDR < 5%, which can be an indication of low quality ChIP (Landt et al., 2012). On the other hand for the PR dataset two datasets in T47D were available (Clarke and Graham, 2012; Yin et al., 2012), we did not employ the datasets published by Yin et al. (Yin et al., 2012) because the experiment was performed with an antiprogesterone (RU486) treatment, which would not be expected to elicit the same binding pattern as PR agonist, and lacked any control sample. MACS distributes read tags from the control sample along the genome to model Poisson distribution, and false discovery rate (FDR) is calculated by swapping control and ChIP samples. Therefore it is recommended for ChIP-seq studies to have an appropriate input control sample (Wilbanks and Facciotti, 2010). ENCODE guidelines also emphasize the importance of using a suitable control dataset to adjust for variable DNA fragment lengths (Landt et al., 2012). Welboren *et al.* (Welboren et al., 2009) studied effects of tamoxifen and fulvestrant treatment on the binding of ER α in MCF7 cell-line, however, due to different cell line this data cannot be compared with datasets available for T47D cell line.

There is a slight difference in the reported low-significance motifs for PR data between this report and the Clarke and Graham study (Clarke and Graham, 2012). The two most significant motifs (PRE and FOXA1) are the same in the two studies, However, Clarke and

Graham found an NF1 half-site as one of the significant motifs and AP-1 sites as non-significant while in this study we found an AP-2 motif higher in significance than the NF1 motif (not shown). This minor difference is due to the difference in binding regions as Clarke and Graham published 6,312 PR bound regions in T47D cells by aligning to the hg18 build of human genome and using the ERANGE peak caller, however, in this study we reported 22,152 PR regions by aligning to the hg19 assembly and using MACS as our peak caller. The two tools have different statistical algorithms that assign significance to the peaks. ERANGE (Enhanced Read Analysis of Gene Expression) algorithm is used for analysis both RNA-Seq and ChIP-Seq data while MACS is exclusively designed to call peaks for ChIP-Seq data (Feng et al., 2011; Mortazavi et al., 2008).

The ER α -PR data were collected from two separate publications where the binding of each factor was studied by stimulation of T-47D cells with estrogen or progesterone independently. Therefore the focus of this study was to examine the correlation of ER α -PR binding patterns which revealed an interesting convergence on specific loci. We studied the association between common regions and nearby genes and found biologically relevant gene pathways. The Myc oncogene, which was most highly associated with binding sites common to ER α and PR, was up-regulated in the ER α regulated gene expression dataset. Myc is a known target of both estrogen and progesterone and plays a key role in the normal breast and breast cancer (Hynes and Stoezle, 2009; Curtis et al., 2012) PR itself is also regulated by both hormones and the PGR gene was highly associated with shared ER α and PR binding regions. Transcriptional regulation by estrogen and progesterone co-treatment in this cell model was not available, however it would be interesting to study the binding of the two factors under the influence of both stimuli (estrogen and progesterone) to observe the impact of converging ER α and PR regulation in comparison to individual stimulation.

In summary, we have evidence for a biologically relevant interplay between PR and ER α in a subset of binding sites in breast cancer cells. Our analysis demonstrated the utility of our developed and published software BiSA (Khushi et al., 2014), which has a comprehensive knowledge base, consisting of transcription factor binding sites and histone modifications collected from previously published studies. Using BiSA we identified that ER α and PR co-locate on a subset of binding sites. The BiSA statistical testing of overlap revealed a low overlap correlation value (OCV) suggesting that the two factors are not obligate cofactors. However, spatial correlation testing using Monte Carlo simulation by BITS, Relevant Distance, Absolute Distance, Jaccard and Projection tests by Genometricorr revealed a statistically significant correlation between the two factors. The ER α , FOXA1, PR, AP-2 and TEAD4 binding motifs are significantly enriched in regions that are bound by both ER α and PR. In addition, gene expression analysis revealed apoptosis as one of the significant biological process by the set of transcripts that were regulated by ER α -PR common region suggesting that their overlap is biologically relevant.

Chapter 6: Discussion

At the time of initial inception of the project in 2009, there was a lack of tools available to explore genomic regions. UCSC Genome Browser (Kent et al., 2002b) and Galaxy provided very limited functions to find overlapping regions. BedTools (Quinlan and Hall, 2010a) was published in 2010, and in later years a number of other tools were published (Tsirigos et al., 2012; Neph et al., 2012; Renaud et al., 2011; Dale et al., 2011a). However, there was no integrated tool available that had pre-loaded data from previous publications, gene annotations and options to visualise the degree of overlap. These were basic needs in most ChIP-Seq studies, therefore, we decided to develop an integrated tool that could perform the above necessary operations and with up-to-date datasets from published studies for comparison of data with previous reports.

In Chapter 2, I surveyed SQL and No-SQL based databases and decided to design the software using SQL-databases because genomic data follows a strict format which fulfils the requirement of SQL-based databases. No-SQL databases are mainly designed to incorporate non-standard data. Moreover, SQL databases provide a powerful query interface and rational design that makes it easy to link data between different tables. To review SQL databases in depth I developed a novel algorithm RegMap (Region Mapping) natively written in SQL to find overlapping or nearby genomic regions. Using the RegMap algorithm I performed performance benchmarking for widely used databases. The benchmarking results for PostgreSQL and MySQL databases were published (Khushi 2015), however, benchmarking for other proprietary databases was not published because of their licensing agreements. Database benchmarking revealed that searching and retrieving information were very efficient for all databases. Results were retrieved in a few seconds (<5s) while searching millions of records. However performance related to basic mathematical calculations such as subtraction greatly varied among the different databases

when large data was processed. Benchmarking identified that PostgreSQL extracts overlapping regions much faster than MySQL, in addition, insertion and data uploads in PostgreSQL were also time-efficient. The RegMap algorithm performed better than database built-in Geo-mapping functions to report intersecting genomic regions. Other limitations of built-in Geo functions are that they cannot be used to report nearby genomic regions easily. The ability to easily identify nearby genomic regions is built into the RegMap algorithm, therefore, with some modifications the algorithm can also be used in other computer science applications where nearby spatial data required to be identify such as finding nearby geometrical lines or streets. RegMap algorithm was implemented in the development of BiSA (Binding Sites Analyser) (Khushi et al., 2014) software using Microsoft SQL Server on Windows version and PostgreSQL on Unix/Linux. Most tools that are available to identify overlapping or non-overlapping regions were written for Unix/Linux environment. On Windows, which is one of the most popular operating systems, there is no comprehensive tool available. Cisgenome (Ji et al., 2011) for Windows has very limited options to compare genomic regions. Cisgenome lacks options to extract nearby regions and restrict on distance from centre of regions. Therefore BiSA for Windows addressed the need for a comprehensive Windows-based application to analyse genomic regions.

Chapter 2 described the BiSA database schema. The database design allows archiving unlimited numbers of genomic regions. In recent years due to a steep reduction in sequencing costs the rate of generation of genomic data has been growing exponentially. Therefore the BiSA database architecture allows researchers to archive and analyse unlimited numbers of genomic regions. The BiSA Windows based Graphical user Interface (GUI) is very easy to operate, however, we identified that installation of the Microsoft SQL Server could be a non-trivial task for biologists. Therefore we developed a web-based

version that runs under Galaxy which could be installed once for unlimited number of users. BiSA has been populated with >1000 datasets of transcription factors and histone modifications. Each dataset has information about genomic assembly, first author, cell line, factor, treatment or other useful information which are saved in notes. Links to publications and raw data are also recorded, which makes it easy to read about the background to the study design.

Chapter 3 describes all options and application of BiSA. BiSA is an integrated graphical user interface (GUI) tool that provides a number of options to study overlapping/non-overlapping or nearby regions, which otherwise is achieved by a number of different tools. Users can visualize genomic overlap results as Venn diagrams and can save chart images for use in publications. BiSA can identify genes associated with binding regions of interest and also the statistical significance of overlapping regions. In BiSA, the statistical significance of overlapping regions is calculated by the IntervalStats (Chikina and Troyanskaya, 2012) tool as this helps in identifying partner factors. BiSA calculates an Overlap Correlation Value (OCV) which is a summary statistic of p-values of overlapping regions less than a value defined as BiSA threshold. We set this threshold to 0.05, however, this can be changed in the BiSA configuration file. An OCV greater than 0.5 is considered a significant correlation between two factors. Other statistical tools such as Genometricorr (Favorov et al., 2012) and BITS (Layer et al., 2013) are designed to identify the statistical significance of the special relationship of two sets of regions.

Chapter 4 describes the utility of the BiSA knowledge base, which provides a great opportunity to mine genomic regions for the identification of biologically relevant relationships. In this chapter, I identified that MACS and HOMER were two most popular peak-caller tools. As I explained in Chapter 1, different peak-callers identify different number of peaks (genomic regions) which could affect calculation of OCV for two

datasets. Therefore, I systematically performed validation for calculating the OCV and identified that this technique can be applied to datasets that were generated using these two peak-callers. Using the data from three different studies I identified that HNF4G nuclear receptor significantly collocated with STAG1 and H3K4me3 promoter marks in HepG2 cell line. STAG1 (Stromal Antigen 1), also known as SA1, is one of the four subunits of the cohesin complex involved in transcription regulation, DNA repair, chromosome condensation, homolog pairing. Therefore, the statistical significant overlap of HNF4G with STAG1 indicates an important underlying biology which could be further explored in the laboratory.

The BiSA knowledge base has a number of datasets describing genomic locations of various factors such as ER α , PR and FOXA1 for the T47D breast cancer cell line. In Chapter 4, I also studied the correlation between these datasets by calculating OCV and drawn a network diagram for significant correlations where OCV was greater than or equal to 0.5. The calculation of OCV is a computationally intense task as each region from one dataset is matched to its closest region in other dataset then the significance of two regions being overlapping or close to each other is calculated. Therefore I narrowed down my selection to 12 datasets based on a number of criteria explained in Chapter 4. The network map revealed that FOXA1 influenced the binding of seven other factors JUND, P300, ER α , GATA3, PR-RU486, and XBP1. This analysis also revealed that the binding sites that were targeted by PR were different under progesterone and anti-progesterin (mifepristone) treatment. Interestingly ER α had significant OCV with PR when stimulated by anti-progesterin while ER α showed no significant overlap with PR when simulated with progesterin. Mifepristone is used as an abortifacient in the first trimester, as emergency contraception and in low dose as a contraceptive. Therefore I hypothesised that in the presence of mifepristone and estradiol, ER α and PR target similar genomic locations in

regulation of gene expression. This relationship could further be validated in the laboratory.

Using BiSA we identified the degree of overlap for datasets in a pair-wise fashion and looked for biologically relevant overlapping datasets. We identified that ~27% of ER α binding regions overlap with PR binding regions in the T47D breast cancer cell line. Since ER α and PR are known to be major players in regulation and progression of breast cancer, the sharing of more than one quarter of binding sites between these two factors was interesting. In Chapter 5, we studied these datasets in great depth by sourcing raw data from original publications. Using BiSA, we identified that the OCV between ER α and PR was only 0.33 when ER α was selected as query and PR was selected as reference and OCV further reduced to 0.26 when PR was selected as query against the whole genome as domain background. Therefore, the ER α -PR OCV less than 0.5 reflected that the correlation was not statistically significant. Based on this finding we concluded that the two factors do not usually share binding sites and the observed overlap suggests that their activities converge on specific DNA loci. Motif analysis revealed ER α , FOXA1, PR, AP-2 and TEAD4 binding motifs are significantly enriched in common ER α -PR regions and gene expression analysis identified apoptosis as one of the significant biological process by the set of transcripts that were regulated by ER α -PR common regions. Therefore ER α -PR analysis suggests that their overlap is biologically relevant.

In future I have planned to extend the project by designing a unique web interface to address some of the limitation of BiSA for Galaxy. For example, although BiSA for Galaxy provides essential web-based operations, at present it doesn't provide a way to filter tool options based on a logged-in user. This means that every user of a Galaxy tool will be presented with the same set of features, therefore, options cannot be customised based on user identification.

In summary, in this thesis I described the development of a bioinformatics resource and tools (BiSA) to identify genomic regions overlap and statistical significance of the overlap. BiSA can annotate genomic regions and degree of overlap can be visualised as a Venn diagram. The BiSA database contains a comprehensive knowledge base and I have demonstrated how BiSA can be used to study various publicly available datasets. These example analyses showed a great utility of BiSA tools and its built-in knowledge base. Therefore I envision that the BiSA resource and tools will continue to support growth and knowledge in bioinformatics and genomic research. I have published three peer-reviewed articles which are attached as appendix to the thesis.

References

- ABDEL-HAFIZ, H. A. & HORWITZ, K. B. 2014. Post-translational modifications of the progesterone receptors. *J Steroid Biochem Mol Biol*, 140, 80-9.
- ACCETTA, M., BARON, R., BOLOSKY, W., GOLUB, D., RASHID, R., TEVANIAN, A. & YOUNG, M. 1986. Mach: A new kernel foundation for UNIX development.
- ADOMAS, A. B., GRIMM, S. A., MALONE, C., TAKAKU, M., SIMS, J. K. & WADE, P. A. 2014. Breast tumor specific mutation in GATA3 affects physiological mechanisms regulating transcription factor turnover. *BMC Cancer*, 14, 278.
- ALBERTS, B. 2008. *Molecular Biology of the Cell: Reference Edition*, Garland Science.
- ALBUISSON, J., ISIDOR, B., GIRAUD, M., PICHON, O., MARSAUD, T., DAVID, A., LE CAIGNEC, C. & BEZIEAU, S. 2011. Identification of two novel mutations in Shh long-range regulator associated with familial pre-axial polydactyly. *Clin Genet*, 79, 371-7.
- ANIBA, M. R., POCH, O. & THOMPSON, J. D. 2010. Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. *Nucleic Acids Res*, 38, 7353-63.
- ARBIZA, L., GRONAU, I., AKSOY, B. A., HUBISZ, M. J., GULKO, B., KEINAN, A. & SIEPEL, A. 2013. Genome-wide inference of natural selection on human transcription factor binding sites. *Nat Genet*, 45, 723-9.
- ARSLAN, A. & YILMAZEL, O. A comparison of relational databases and information retrieval libraries on turkish text retrieval. Natural Language Processing and Knowledge Engineering, 2008. NLP-KE'08. International Conference on, 2008. IEEE, 1-8.

- ARVEY, A., AGIUS, P., NOBLE, W. S. & LESLIE, C. 2012. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res*, 22, 1723-34.
- ASZODI, A. 2012. MULTOVL: fast multiple overlaps of genomic regions. *Bioinformatics*, 28, 3318-9.
- AUGELLO, M. A., HICKEY, T. E. & KNUDSEN, K. E. 2011. FOXA1: master of steroid receptor function in cancer. *EMBO J*, 30, 3885-94.
- AZAD, N., ZAHNOW, C. A., RUDIN, C. M. & BAYLIN, S. B. 2013. The future of epigenetic therapy in solid tumours--lessons from the past. *Nat Rev Clin Oncol*, 10, 256-66.
- AZUARA, V., PERRY, P., SAUER, S., SPIVAKOV, M., JORGENSEN, H. F., JOHN, R. M., GOUTI, M., CASANOVA, M., WARNES, G., MERKENSCHLAGER, M. & FISHER, A. G. 2006. Chromatin signatures of pluripotent cell lines. *Nat Cell Biol*, 8, 532-8.
- BAILEY, T., KRAJEWSKI, P., LADUNGA, I., LEFEBVRE, C., LI, Q., LIU, T., MADRIGAL, P., TASLIM, C. & ZHANG, J. 2013. Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput Biol*, 9, e1003326.
- BAILEY, T. L. & MACHANICK, P. 2012. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res*, 40, e128.
- BALLARE, C., CASTELLANO, G., GAVEGLIA, L., ALTHAMMER, S., GONZALEZ-VALLINAS, J., EYRAS, E., LE DILY, F., ZAURIN, R., SORONELLAS, D., VICENT, G. P. & BEATO, M. 2013. Nucleosome-driven transcription factor binding and gene regulation. *Mol Cell*, 49, 67-79.
- BALLARÉ, C., CASTELLANO, G., GAVEGLIA, L., ALTHAMMER, S., GONZÁLEZ-VALLINAS, J., EYRAS, E., LE DILY, F., ZAURIN, R., SORONELLAS, D.,

- VICENT, GUILLERMO P. & BEATO, M. 2013. Nucleosome-Driven Transcription Factor Binding and Gene Regulation. *Molecular Cell*, 49, 67-79.
- BANERJI, J., RUSCONI, S. & SCHAFFNER, W. 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27, 299-308.
- BARASH, Y., VAQUERO-GARCIA, J., GONZALEZ-VALLINAS, J., XIONG, H. Y., GAO, W., LEE, L. J. & FREY, B. J. 2013. AVISPA: a web tool for the prediction and analysis of alternative splicing. *Genome Biol*, 14, R114.
- BARSKI, A. & ZHAO, K. 2009. Genomic location analysis by ChIP-Seq. *J Cell Biochem*, 107, 11-8.
- BAUER, U. M., DAUJAT, S., NIELSEN, S. J., NIGHTINGALE, K. & KOUZARIDES, T. 2002. Methylation at arginine 17 of histone H3 is linked to gene activation. *EMBO Rep*, 3, 39-44.
- BAYLIN, S. B. & HERMAN, J. G. 2000. DNA hypermethylation in tumorigenesis: epigenetics joins genetics. *Trends Genet*, 16, 168-74.
- BEJERANO, G., PHEASANT, M., MAKUNIN, I., STEPHEN, S., KENT, W. J., MATTICK, J. S. & HAUSSLER, D. 2004. Ultraconserved elements in the human genome. *Science*, 304, 1321-5.
- BELJANSKI, M. 2013. The Regulation of DNA Replication and Transcription. New York: Demos Medical Publishing.
- BENAGIANO, G., BASTIANELLI, C. & FARRIS, M. 2008. Selective progesterone receptor modulators 2: use in reproductive medicine. *Expert Opin Pharmacother*, 9, 2473-85.
- BENAGIANO, G., BASTIANELLI, C., FARRIS, M. & BROSENS, I. 2014. Selective progesterone receptor modulators: an update. *Expert Opin Pharmacother*, 15, 1403-15.

- BERG, T. 2011. Small-molecule modulators of c-Myc/Max and Max/Max interactions. *Curr Top Microbiol Immunol*, 348, 139-49.
- BERMAN, B. P., WEISENBERGER, D. J., AMAN, J. F., HINOUE, T., RAMJAN, Z., LIU, Y., NOUSHMEHR, H., LANGE, C. P., VAN DIJK, C. M., TOLLENAAR, R. A., VAN DEN BERG, D. & LAIRD, P. W. 2012. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat Genet*, 44, 40-6.
- BERNARDO, G. M. & KERI, R. A. 2012. FOXA1: a transcription factor with parallel functions in development and cancer. *Biosci Rep*, 32, 113-30.
- BERNSTEIN, B. E., MIKKELSEN, T. S., XIE, X., KAMAL, M., HUEBERT, D. J., CUFF, J., FRY, B., MEISSNER, A., WERNIG, M., PLATH, K., JAENISCH, R., WAGSCHAL, A., FEIL, R., SCHREIBER, S. L. & LANDER, E. S. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125, 315-26.
- BERNSTEIN, P. A., HADZILACOS, V. & GOODMAN, N. 1987. *Concurrency control and recovery in database systems*, Addison-wesley New York, 370.
- BLANKENBERG, D., GORDON, A., VON KUSTER, G., CORAOR, N., TAYLOR, J., NEKRUTENKO, A. & GALAXY, T. 2010. Manipulation of FASTQ data with Galaxy. *Bioinformatics*, 26, 1783-5.
- BOSE, S., MISHRA, P., SETHURAMAN, P. & TAHERI, R. 2009. Benchmarking Database Performance in a Virtual Environment. *In*: NAMBIAR, R. & POESS, M. (eds.) *Performance Evaluation and Benchmarking*. Springer Berlin Heidelberg.
- BOYLE, A. P., GUINNEY, J., CRAWFORD, G. E. & FUREY, T. S. 2008. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, 24, 2537-8.

- BRAY, T., PAOLI, J., SPERBERG-MCQUEEN, C. M., MALER, E. & YERGEAU, F. 1997. Extensible markup language (XML). *World Wide Web Journal*, 2, 27-66.
- BRENOWITZ, M., SENEAR, D. F., SHEA, M. A. & ACKERS, G. K. 1986. Quantitative DNase footprint titration: a method for studying protein-DNA interactions. *Methods Enzymol*, 130, 132-81.
- BUECKER, C. & WYSOCKA, J. 2012. Enhancers as information integration hubs in development: lessons from genomics. *Trends Genet*, 28, 276-84.
- BULGER, M. & GROUDINE, M. 2011a. Functional and mechanistic diversity of distal transcription enhancers. *Cell*, 144, 327-39.
- BULGER, M. & GROUDINE, M. 2011b. Functional and Mechanistic Diversity of Distal Transcription Enhancers. *Cell*, 144, 327-339.
- BULUN, S. E. 2014. Aromatase and estrogen receptor alpha deficiency. *Fertil Steril*, 101, 323-9.
- CADDOO, K. A., FORNIER, M. N. & MORRIS, P. G. 2013. Biological subtypes of breast cancer: current concepts and implications for recurrence patterns. *Q J Nucl Med Mol Imaging*, 57, 312-21.
- CALO, E. & WYSOCKA, J. 2013. Modification of enhancer chromatin: what, how, and why? *Mol Cell*, 49, 825-37.
- CAREY, M. & SMALE, S. T. 2001. *Transcriptional Regulation in Eukaryotes: Concepts, Strategies, and Techniques*, Cold Spring Harbor Laboratory Press.
- CARLBERG, C. 2014. Genome-wide (over)view on the actions of vitamin D. *Front Physiol*, 5, 167.
- CARROLL, J. S., LIU, X. S., BRODSKY, A. S., LI, W., MEYER, C. A., SZARY, A. J., EECKHOUTE, J., SHAO, W., HESTERMANN, E. V., GEISTLINGER, T. R., FOX, E. A., SILVER, P. A. & BROWN, M. 2005. Chromosome-wide mapping of

estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell*, 122, 33-43.

- CARROLL, J. S., MEYER, C. A., SONG, J., LI, W., GEISTLINGER, T. R., EECKHOUTE, J., BRODSKY, A. S., KEETON, E. K., FERTUCK, K. C., HALL, G. F., WANG, Q., BEKIRANOV, S., SEMENTCHENKO, V., FOX, E. A., SILVER, P. A., GINGERAS, T. R., LIU, X. S. & BROWN, M. 2006. Genome-wide analysis of estrogen receptor binding sites. *Nat Genet*, 38, 1289-97.
- CATTELL, R. 2011. Scalable SQL and NoSQL data stores. *ACM SIGMOD Record*, 39, 12-27.
- CERAMI, E. G., BADER, G. D., GROSS, B. E. & SANDER, C. 2006. cPath: open source software for collecting, storing, and querying biological pathways. *BMC Bioinformatics*, 7, 497.
- CHABBERT-BUFFET, N., ESBER, N. & BOUCHARD, P. 2014. Fibroid growth and medical options for treatment. *Fertil Steril*, 102, 630-9.
- CHALBOS, D., VIGNON, F., KEYDAR, I. & ROCHEFORT, H. 1982. Estrogens stimulate cell proliferation and induce secretory proteins in a human breast cancer cell line (T47D). *J Clin Endocrinol Metab*, 55, 276-83.
- CHANG, F., DEAN, J., GHEMAWAT, S., HSIEH, W. C., WALLACH, D. A., BURROWS, M., CHANDRA, T., FIKES, A. & GRUBER, R. E. 2008. Bigtable: A distributed storage system for structured data. *Acm Transactions on Computer Systems*, 26, 4.
- CHARTS, G. 2014. *Google Charts* [Online]. Available: <http://code.google.com/apis/chart/>.
- CHEN, P., TONG, X. L., LI, D. D., FU, M. Y., HE, S. Z., HU, H., XIANG, Z. H., LU, C. & DAI, F. Y. 2013. Antennapedia is involved in the development of thoracic legs and segmentation in the silkworm, *Bombyx mori*. *Heredity (Edinb)*, 111, 182-8.

- CHEN, X., ILIOPOULOS, D., ZHANG, Q., TANG, Q., GREENBLATT, M. B.,
HATZIAPOSTOULOU, M., LIM, E., TAM, W. L., NI, M., CHEN, Y., MAI, J.,
SHEN, H., HU, D. Z., ADORO, S., HU, B., SONG, M., TAN, C., LANDIS, M. D.,
FERRARI, M., SHIN, S. J., BROWN, M., CHANG, J. C., LIU, X. S. &
GLIMCHER, L. H. 2014. XBP1 promotes triple-negative breast cancer by
controlling the HIF1alpha pathway. *Nature*, 508, 103-7.
- CHEN, X., XU, H., YUAN, P., FANG, F., HUSS, M., VEGA, V. B., WONG, E., ORLOV,
Y. L., ZHANG, W., JIANG, J., LOH, Y. H., YEO, H. C., YEO, Z. X., NARANG,
V., GOVINDARAJAN, K. R., LEONG, B., SHAHAB, A., RUAN, Y.,
BOURQUE, G., SUNG, W. K., CLARKE, N. D., WEI, C. L. & NG, H. H. 2008.
Integration of external signaling pathways with the core transcriptional network in
embryonic stem cells. *Cell*, 133, 1106-17.
- CHEUNG, E. & KRAUS, W. L. 2010. Genomic Analyses of Hormone Signaling and Gene
Regulation. *Annual Review of Physiology*.
- CHIKINA, M. D. & TROYANSKAYA, O. G. 2012. An effective statistical evaluation of
ChIPseq dataset similarity. *Bioinformatics*, 28, 607-13.
- CLAESSENS, F. & GEWIRTH, D. T. 2004. DNA recognition by nuclear receptors.
Essays Biochem, 40, 59-72.
- CLARKE, C. L. & GRAHAM, J. D. 2012. Non-overlapping progesterone receptor
cistromes contribute to cell-specific transcriptional outcomes. *PLoS One*, 7,
e35859.
- COCK, P. J., ANTAO, T., CHANG, J. T., CHAPMAN, B. A., COX, C. J., DALKE, A.,
FRIEDBERG, I., HAMELRYCK, T., KAUFF, F., WILCZYNSKI, B. & DE
HOON, M. J. 2009. Biopython: freely available Python tools for computational
molecular biology and bioinformatics. *Bioinformatics*, 25, 1422-3.

- COCK, P. J., FIELDS, C. J., GOTO, N., HEUER, M. L. & RICE, P. M. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res*, 38, 1767-71.
- COCK, P. J., GRUNING, B. A., PASZKIEWICZ, K. & PRITCHARD, L. 2013. Galaxy tools and workflows for sequence analysis with applications in molecular plant pathology. *PeerJ*, 1, e167.
- COLLAS, P. & DAHL, J. A. 2008. Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation. *Front Biosci*, 13, 929-43.
- CONRAD, T. 2006. PostgreSQL vs. MySQL vs. Commercial Databases: It's All About What You Need. Technical report, Devx, 2004. cosmoglobecorp.com/pdf/PostgreSQL_vs_MySQL_vs_DB2_vs_MSSQL_vs_Oracle.pdf.
- CONSALVI, S., SACCONI, V., GIORDANI, L., MINETTI, G., MOZZETTA, C. & PURI, P. L. 2011. Histone deacetylase inhibitors in the treatment of muscular dystrophies: epigenetic drugs for genetic diseases. *Mol Med*, 17, 457-65.
- COOPER, G. M. 2000. *Eukaryotic RNA Polymerases and General Transcription Factors*, Sunderland (MA): Sinauer Associates, 2nd edition.
- COPELAND, R. 2010. *Essential sqlalchemy*, O'Reilly.
- CREYGHTON, M. P., CHENG, A. W., WELSTEAD, G. G., KOOISTRA, T., CAREY, B. W., STEINE, E. J., HANNA, J., LODATO, M. A., FRAMPTON, G. M., SHARP, P. A., BOYER, L. A., YOUNG, R. A. & JAENISCH, R. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A*, 107, 21931-6.
- CROOKS, G. E., HON, G., CHANDONIA, J. M. & BRENNER, S. E. 2004. WebLogo: a sequence logo generator. *Genome Res*, 14, 1188-90.
- CURTIS, C., SHAH, S. P., CHIN, S. F., TURASHVILI, G., RUEDA, O. M., DUNNING, M. J., SPEED, D., LYNCH, A. G., SAMARAJIWA, S., YUAN, Y., GRAF, S.,

- HA, G., HAFFARI, G., BASHASHATI, A., RUSSELL, R., MCKINNEY, S., GROUP, M., LANGEROD, A., GREEN, A., PROVENZANO, E., WISHART, G., PINDER, S., WATSON, P., MARKOWETZ, F., MURPHY, L., ELLIS, I., PURUSHOTHAM, A., BORRESEN-DALE, A. L., BRENTON, J. D., TAVARE, S., CALDAS, C. & APARICIO, S. 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486, 346-52.
- CYR, A. R., KULAK, M. V., PARK, J. M., BOGACHEK, M. V., SPANHEIMER, P. M., WOODFIELD, G. W., WHITE-BAER, L. S., O'MALLEY, Y. Q., SUGG, S. L., OLIVIER, A. K., ZHANG, W., DOMANN, F. E. & WEIGEL, R. J. 2015. TFAP2C governs the luminal epithelial phenotype in mammary development and carcinogenesis. *Oncogene*, 34, 436-44.
- D'ABREO, N. & HINDENBURG, A. A. 2013. Sex hormone receptors in breast cancer. *Vitam Horm*, 93, 99-133.
- DALE, R. K., PEDERSEN, B. S. & QUINLAN, A. R. 2011a. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics*, 27, 3423-4.
- DALE, R. K., PEDERSEN, B. S. & QUINLAN, A. R. 2011b. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics*.
- DANG, C. V. 2012. MYC on the path to cancer. *Cell*, 149, 22-35.
- DAS, P. M., RAMACHANDRAN, K., VANWERT, J. & SINGAL, R. 2004. Chromatin immunoprecipitation assay. *Biotechniques*, 37, 961-9.
- DE HOON, M. J., CHAPMAN, B. & FRIEDBERG, I. 2003. Bioinformatics and computational biology with Biopython. *Genome Informatics Series*, 298-299.
- DE WIT, E. & DE LAAT, W. 2012. A decade of 3C technologies: insights into nuclear organization. *Genes Dev*, 26, 11-24.

- DEAN, A. 2011. In the loop: long range chromatin interactions and gene regulation. *Brief Funct Genomics*, 10, 3-10.
- DELGADO, M. D. & LEON, J. 2006. Gene expression regulation and cancer. *Clin Transl Oncol*, 8, 780-7.
- DI GIACOMO, M. 2005. MySQL: lessons learned on a digital library. *Software, IEEE*, 22, 10-13.
- DJEBALI, S., DAVIS, C. A., MERKEL, A., DOBIN, A., LASSMANN, T., MORTAZAVI, A., TANZER, A., LAGARDE, J., LIN, W., SCHLESINGER, F., XUE, C., MARINOV, G. K., KHATUN, J., WILLIAMS, B. A., ZALESKI, C., ROZOWSKY, J., RODER, M., KOKOCINSKI, F., ABDELHAMID, R. F., ALIOTO, T., ANTOSHECHKIN, I., BAER, M. T., BAR, N. S., BATUT, P., BELL, K., BELL, I., CHAKRABORTTY, S., CHEN, X., CHRAST, J., CURADO, J., DERRIEN, T., DRENKOW, J., DUMAIS, E., DUMAIS, J., DUTTAGUPTA, R., FALCONNET, E., FASTUCA, M., FEJES-TOTH, K., FERREIRA, P., FOISSAC, S., FULLWOOD, M. J., GAO, H., GONZALEZ, D., GORDON, A., GUNAWARDENA, H., HOWALD, C., JHA, S., JOHNSON, R., KAPRANOV, P., KING, B., KINGSWOOD, C., LUO, O. J., PARK, E., PERSAUD, K., PREALL, J. B., RIBECA, P., RISK, B., ROBYR, D., SAMMETH, M., SCHAFFER, L., SEE, L. H., SHAHAB, A., SKANCKE, J., SUZUKI, A. M., TAKAHASHI, H., TILGNER, H., TROUT, D., WALTERS, N., WANG, H., WROBEL, J., YU, Y., RUAN, X., HAYASHIZAKI, Y., HARROW, J., GERSTEIN, M., HUBBARD, T., REYMOND, A., ANTONARAKIS, S. E., HANNON, G., GIDDINGS, M. C., RUAN, Y., WOLD, B., CARNINCI, P., GUIGO, R. & GINGERAS, T. R. 2012. Landscape of transcription in human cells. *Nature*, 489, 101-8.
- DUDLEY, J. T. & BUTTE, A. J. 2009. A quick guide for developing effective bioinformatics programming skills. *PLoS Comput Biol*, 5, e1000589.

- ECKLER, M. J., LARKIN, K. A., MCKENNA, W. L., KATZMAN, S., GUO, C., ROQUE, R., VISEL, A., RUBENSTEIN, J. L. & CHEN, B. 2014. Multiple conserved regulatory domains promote Fezf2 expression in the developing cerebral cortex. *Neural Dev*, 9, 6.
- EDGAR, R., DOMRACHEV, M. & LASH, A. E. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30, 207-10.
- EECKHOUTE, J., METIVIER, R. & SALBERT, G. 2009. Defining specificity of transcription factor regulatory activities. *J Cell Sci*, 122, 4027-34.
- EGGER, G., LIANG, G., APARICIO, A. & JONES, P. A. 2004. Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, 429, 457-63.
- ELNITSKI, L., JIN, V. X., FARNHAM, P. J. & JONES, S. J. 2006. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res*, 16, 1455-64.
- ERIC JONES, T. O., PEARU PETERSON AND OTHERS. 2001. *SciPy: Open Source Scientific Tools for Python* [Online]. Available: <http://www.scipy.org/>.
- ESTELLER, M., SILVA, J. M., DOMINGUEZ, G., BONILLA, F., MATIAS-GUIU, X., LERMA, E., BUSSAGLIA, E., PRAT, J., HARKES, I. C., REPASKY, E. A., GABRIELSON, E., SCHUTTE, M., BAYLIN, S. B. & HERMAN, J. G. 2000. Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. *J Natl Cancer Inst*, 92, 564-9.
- FAVOROV, A., MULARONI, L., COPE, L. M., MEDVEDEVA, Y., MIRONOV, A. A., MAKEEV, V. J. & WHEELAN, S. J. 2012. Exploring massive, genome scale datasets with the GenometriCorr package. *PLoS Comput Biol*, 8, e1002529.

- FEJES, A. P., ROBERTSON, G., BILENKY, M., VARHOL, R., BAINBRIDGE, M. & JONES, S. J. 2008. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, 24, 1729-30.
- FENG, J., LIU, T., QIN, B., ZHANG, Y. & LIU, X. S. 2012. Identifying ChIP-seq enrichment using MACS. *Nat Protoc*, 7, 1728-40.
- FENG, J., LIU, T. & ZHANG, Y. 2011. Using MACS to identify peaks from ChIP-Seq data. *Curr Protoc Bioinformatics*, Chapter 2, Unit 2 14.
- FERG, S. 2011. *Python & Java: A Side-by-Side Comparison* [Online]. Available: <http://pythonconquerstheuniverse.wordpress.com/2009/10/03/python-java-a-side-by-side-comparison/>.
- FIORENTINO, F. P. & GIORDANO, A. 2012. The tumor suppressor role of CTCF. *J Cell Physiol*, 227, 479-92.
- FIORITO, E., KATIKA, M. R. & HURTADO, A. 2013. Cooperating transcription factors mediate the function of estrogen receptor. *Chromosoma*, 122, 1-12.
- FLANAGAN, D. & MATSUMOTO, Y. 2008. *The ruby programming language*, O'Reilly.
- FORTIER, P. J., WOLFE, V. F. & PRICHARD, J. J. Flexible real-time SQL transactions. *Real-Time Systems Symposium, 1994., Proceedings., 1994. IEEE*, 276-280.
- FRIEDMAN, J. R. & KAESTNER, K. H. 2006. The Foxa family of transcription factors in development and metabolism. *Cell Mol Life Sci*, 63, 2317-28.
- FRIETZE, S., LAN, X., JIN, V. X. & FARNHAM, P. J. 2010. Genomic targets of the KRAB and SCAN domain-containing zinc finger protein 263. *J Biol Chem*, 285, 1393-403.
- FUREY, T. S. 2012. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet*, 13, 840-52.
- GADALETA, R. M. & MAGNANI, L. 2014. Nuclear receptors and chromatin: an inducible couple. *J Mol Endocrinol*, 52, R137-49.

- GALAS, D. J. & SCHMITZ, A. 1978. DNase footprinting a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Research*, 5, 3157-3170.
- GALAXYTEAM 2013. Galaxy Tool XML File.
- GAVRILOV, A., EIVAZOVA, E., PRIOZHKOVA, I., LIPINSKI, M., RAZIN, S. & VASSETZKY, Y. 2009. Chromosome conformation capture (from 3C to 5C) and its CHIP-based modification. *Methods Mol Biol*, 567, 171-88.
- GERSTEIN, M. B., BRUCE, C., ROZOWSKY, J. S., ZHENG, D., DU, J., KORBEL, J. O., EMANUELSSON, O., ZHANG, Z. D., WEISSMAN, S. & SNYDER, M. 2007. What is a gene, post-ENCODE? History and updated definition. *Genome Res*, 17, 669-81.
- GERSTEIN, M. B., KUNDAJE, A., HARIHARAN, M., LANDT, S. G., YAN, K. K., CHENG, C., MU, X. J., KHURANA, E., ROZOWSKY, J., ALEXANDER, R., MIN, R., ALVES, P., ABYZOV, A., ADDLEMAN, N., BHARDWAJ, N., BOYLE, A. P., CAYTING, P., CHAROS, A., CHEN, D. Z., CHENG, Y., CLARKE, D., EASTMAN, C., EUSKIRCHEN, G., FRIETZE, S., FU, Y., GERTZ, J., GRUBERT, F., HARMANCI, A., JAIN, P., KASOWSKI, M., LACROUTE, P., LENG, J., LIAN, J., MONAHAN, H., O'GEEN, H., OUYANG, Z., PARTRIDGE, E. C., PATACSIL, D., PAULI, F., RAHA, D., RAMIREZ, L., REDDY, T. E., REED, B., SHI, M., SLIFER, T., WANG, J., WU, L., YANG, X., YIP, K. Y., ZILBERMAN-SCHAPIRA, G., BATZOGLOU, S., SIDOW, A., FARNHAM, P. J., MYERS, R. M., WEISSMAN, S. M. & SNYDER, M. 2012. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489, 91-100.
- GERTZ, J., REDDY, T. E., VARLEY, K. E., GARABEDIAN, M. J. & MYERS, R. M. 2012. Genistein and bisphenol A exposure cause estrogen receptor 1 to bind thousands of sites in a cell type-specific manner. *Genome Res*, 22, 2153-62.

- GERTZ, J., SAVIC, D., VARLEY, K. E., PARTRIDGE, E. C., SAFI, A., JAIN, P., COOPER, G. M., REDDY, T. E., CRAWFORD, G. E. & MYERS, R. M. 2013. Distinct properties of cell-type-specific and shared transcription factor binding sites. *Mol Cell*, 52, 25-36.
- GHOSH, A. K. & VARGA, J. 2007. The transcriptional coactivator and acetyltransferase p300 in fibroblast biology and fibrosis. *J Cell Physiol*, 213, 663-71.
- GIARDINE, B., RIEMER, C., HARDISON, R. C., BURHANS, R., ELNITSKI, L., SHAH, P., ZHANG, Y., BLANKENBERG, D., ALBERT, I., TAYLOR, J., MILLER, W., KENT, W. J. & NEKRUTENKO, A. 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*, 15, 1451-5.
- GILFILLAN, G. D., HUGHES, T., SHENG, Y., HJORTH AUG, H. S., STRAUB, T., GERVIN, K., HARRIS, J. R., UNDLIEN, D. E. & LYLE, R. 2012. Limitations and possibilities of low cell number ChIP-seq. *BMC Genomics*, 13, 645.
- GODING, C. R. & O'HARE, P. 1989. Herpes simplex virus Vmw65-octamer binding protein interaction: a paradigm for combinatorial control of transcription. *Virology*, 173, 363-7.
- GOECKS, J., NEKRUTENKO, A., TAYLOR, J. & GALAXY, T. 2010a. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11, R86.
- GOECKS, J., NEKRUTENKO, A., TAYLOR, J. & TEAM, T. G. 2010b. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11, R86.
- GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLER, H., LOH, M. L., DOWNING, J. R., CALIGIURI, M. A., BLOOMFIELD, C. D. & LANDER, E. S. 1999. Molecular classification of

cancer: class discovery and class prediction by gene expression monitoring.

Science, 286, 531-7.

GOTO, N., PRINS, P., NAKAO, M., BONNAL, R., AERTS, J. & KATAYAMA, T. 2010.

BioRuby: bioinformatics software for the Ruby programming language.

Bioinformatics, 26, 2617-9.

GOTO, Y., GOMEZ, M., BROCKDORFF, N. & FEIL, R. 2002. Differential patterns of

histone methylation and acetylation distinguish active and repressed alleles at X-

linked genes. *Cytogenetic and Genome Research*, 99, 66-74.

GRANT, C. E., BAILEY, T. L. & NOBLE, W. S. 2011. FIMO: scanning for occurrences

of a given motif. *Bioinformatics*, 27, 1017-8.

GROBER, O. M., MUTARELLI, M., GIURATO, G., RAVO, M., CICATIELLO, L., DE

FILIPPO, M. R., FERRARO, L., NASSA, G., PAPA, M. F., PARIS, O.,

TARALLO, R., LUO, S., SCHROTH, G. P., BENES, V. & WEISZ, A. 2011.

Global analysis of estrogen receptor beta binding to breast cancer cell genome

reveals an extensive interplay with estrogen receptor alpha for target gene

regulation. *BMC Genomics*, 12, 36.

GU, F., HSU, H. K., HSU, P. Y., WU, J., MA, Y., PARVIN, J., HUANG, T. H. & JIN, V.

X. 2010. Inference of hierarchical regulatory network of estrogen-dependent breast

cancer through ChIP-based data. *BMC Syst Biol*, 4, 170.

GUPTA, S., STAMATOYANNOPOULOS, J. A., BAILEY, T. L. & NOBLE, W. S. 2007.

Quantifying similarity between motifs. *Genome Biol*, 8, R24.

HAERDER, T. & REUTER, A. 1983. Principles of Transaction-Oriented Database

Recovery. *Computing Surveys*, 15, 287-317.

HAMPSHIRE, A. J., RUSLING, D. A., BROUGHTON-HEAD, V. J. & FOX, K. R. 2007.

Footprinting: a method for determining the sequence selectivity, affinity and

kinetics of DNA-binding ligands. *Methods*, 42, 128-40.

- HAVE, C. T. & JENSEN, L. J. 2013. Are graph databases ready for bioinformatics?
Bioinformatics, 29, 3107-8.
- HE-QIN, Z. 2007. Best scheme of design dynamic website: Apache+ PHP+ MySQL [J].
Computer Engineering and Design, 4, 058.
- HE, C., WANG, X. & ZHANG, M. Q. 2014. Nucleosome eviction and multiple co-factor binding predict estrogen-receptor-alpha-associated long-range interactions. *Nucleic Acids Res*, 42, 6935-44.
- HEINZ, S., BENNER, C., SPANN, N., BERTOLINO, E., LIN, Y. C., LASLO, P., CHENG, J. X., MURRE, C., SINGH, H. & GLASS, C. K. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, 38, 576-89.
- HERKE, S. W., SERIO, N. V. & ROGERS, B. T. 2005. Functional analyses of tiptop and antennapedia in the embryonic development of *Oncopeltus fasciatus* suggests an evolutionary pathway from ground state to insect legs. *Development*, 132, 27-34.
- HERMIDA, L., POUSSIN, C., STADLER, M. B., GUBIAN, S., SEWER, A., GAIDATZIS, D., HOTZ, H. R., MARTIN, F., BELCASTRO, V., CANO, S., PEITSCH, M. C. & HOENG, J. 2013. Conifero: an integrated contrast data and gene set platform for computational analysis and biological interpretation of omics data. *BMC Genomics*, 14, 514.
- HEROLD, M., BARTKUHN, M. & RENKAWITZ, R. 2012. CTCF: insights into insulator function during development. *Development*, 139, 1045-57.
- HERTZ, G. Z. & STORMO, G. D. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15, 563-77.
- HIPP, D. R., KENNEDY, D. & MISTACHKIN, J. 2013. *The SQLite Development Team* [Online]. Available: <http://www.sqlite.org/crew.html>.

- HO, J., BISHOP, E., KARCHENKO, P., NEGRE, N., WHITE, K. & PARK, P. 2011. ChIP-chip versus ChIP-seq: Lessons for experimental design and data analysis. *BMC Genomics*, 12, 134.
- HOLLAND, R. C., DOWN, T. A., POCOCK, M., PRLIC, A., HUEN, D., JAMES, K., FOISY, S., DRAGER, A., YATES, A., HEUER, M. & SCHREIBER, M. J. 2008. BioJava: an open-source framework for bioinformatics. *Bioinformatics*, 24, 2096-7.
- HOLLIDAY, R. 2006. Epigenetics: a historical overview. *Epigenetics*, 1, 76-80.
- HOLWERDA, S. J. & DE LAAT, W. 2013. CTCF: the protein, the binding partners, the binding sites and their chromatin loops. *Philos Trans R Soc Lond B Biol Sci*, 368, 20120369.
- HOOPER, C. & HILLIKER, A. 2013. Packing them up and dusting them off: RNA helicases and mRNA storage. *Biochim Biophys Acta*, 1829, 824-34.
- HU, M., YU, J., TAYLOR, J. M., CHINNAIYAN, A. M. & QIN, Z. S. 2010. On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res*, 38, 2154-67.
- HU, Q., LUO, Z., XU, T., ZHANG, J. Y., ZHU, Y., CHEN, W. X., ZHONG, S. L., ZHAO, J. H. & TANG, J. H. 2014. FOXA1: a promising prognostic marker in breast cancer. *Asian Pac J Cancer Prev*, 15, 11-6.
- HUA, S., KITTLER, R. & WHITE, K. P. 2009. Genomic antagonism between retinoic acid and estrogen signaling in breast cancer. *Cell*, 137, 1259-71.
- HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4, 44-57.
- HUBBARD, T., BARKER, D., BIRNEY, E., CAMERON, G., CHEN, Y., CLARK, L., COX, T., CUFF, J., CURWEN, V., DOWN, T., DURBIN, R., EYRAS, E., GILBERT, J., HAMMOND, M., HUMINIECKI, L., KASPRZYK, A.,

- LEHVASLAIHO, H., LIJNZAAD, P., MELSOPP, C., MONGIN, E., PETTETT, R., POCOCK, M., POTTER, S., RUST, A., SCHMIDT, E., SEARLE, S., SLATER, G., SMITH, J., SPOONER, W., STABENAU, A., STALKER, J., STUPKA, E., URETA-VIDAL, A., VASTRIK, I. & CLAMP, M. 2002. The Ensembl genome database project. *Nucleic Acids Res*, 30, 38-41.
- HURTADO, A., HOLMES, K. A., GEISTLINGER, T. R., HUTCHESON, I. R., NICHOLSON, R. I., BROWN, M., JIANG, J., HOWAT, W. J., ALI, S. & CARROLL, J. S. 2008. Regulation of ERBB2 by oestrogen receptor-PAX2 determines response to tamoxifen. *Nature*, 456, 663-6.
- HYNES, N. E. & STOELZLE, T. 2009. Key signalling nodes in mammary gland development and cancer: *Myc. Breast Cancer Res*, 11, 210.
- ISHIKAWA, H., ISHI, K., SERNA, V. A., KAKAZU, R., BULUN, S. E. & KURITA, T. 2010. Progesterone is essential for maintenance and growth of uterine leiomyoma. *Endocrinology*, 151, 2433-42.
- JAGOTA, A. 2004. *Perl for Bioinformatics*, Arun Jagota.
- JAMIESON, C., SHARMA, M. & HENDERSON, B. R. 2012. Wnt signaling from membrane to nucleus: beta-catenin caught in a loop. *Int J Biochem Cell Biol*, 44, 847-50.
- JEMAL, A., BRAY, F., CENTER, M. M., FERLAY, J., WARD, E. & FORMAN, D. 2011. Global cancer statistics. *CA Cancer J Clin*, 61, 69-90.
- JI, H., JIANG, H., MA, W., JOHNSON, D. S., MYERS, R. M. & WONG, W. H. 2008. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol*, 26, 1293-300.
- JI, H., JIANG, H., MA, W. & WONG, W. H. 2011. Using CisGenome to analyze ChIP-chip and ChIP-seq data. *Curr Protoc Bioinformatics*, Chapter 2, Unit2 13.

- JIANG, J. & LEVINE, M. 1993. Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen. *Cell*, 72, 741-52.
- JIN, H. J., ZHAO, J. C., WU, L., KIM, J. & YU, J. 2014. Cooperativity and equilibrium with FOXA1 define the androgen receptor transcriptional program. *Nat Commun*, 5, 3972.
- JOHNSON, D. S., LI, W., GORDON, D. B., BHATTACHARJEE, A., CURRY, B., GHOSH, J., BRIZUELA, L., CARROLL, J. S., BROWN, M., FLICEK, P., KOCH, C. M., DUNHAM, I., BIEDA, M., XU, X., FARNHAM, P. J., KAPRANOV, P., NIX, D. A., GINGERAS, T. R., ZHANG, X., HOLSTER, H., JIANG, N., GREEN, R. D., SONG, J. S., MCCUINE, S. A., ANTON, E., NGUYEN, L., TRINKLEIN, N. D., YE, Z., CHING, K., HAWKINS, D., REN, B., SCACHERI, P. C., ROZOWSKY, J., KARPIKOV, A., EUSKIRCHEN, G., WEISSMAN, S., GERSTEIN, M., SNYDER, M., YANG, A., MOQTADERI, Z., HIRSCH, H., SHULHA, H. P., FU, Y., WENG, Z., STRUHL, K., MYERS, R. M., LIEB, J. D. & LIU, X. S. 2008. Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Res*, 18, 393-403.
- JOHNSON, D. S., MORTAZAVI, A., MYERS, R. M. & WOLD, B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316, 1497-502.
- JOHNSTON, S. J. & CARROLL, J. S. 2015. Transcription factors and chromatin proteins as therapeutic targets in cancer. *Biochim Biophys Acta*, 1855, 183-192.
- JOLMA, A., YAN, J., WHITINGTON, T., TOIVONEN, J., NITTA, K. R., RASTAS, P., MORGUNOVA, E., ENGE, M., TAIPALE, M., WEI, G., PALIN, K., VAQUERIZAS, J. M., VINCENTELLI, R., LUSCOMBE, N. M., HUGHES, T. R., LEMAIRE, P., UKKONEN, E., KIVIOJA, T. & TAIPALE, J. 2013. DNA-binding specificities of human transcription factors. *Cell*, 152, 327-39.

- JONES, P. A. & BAYLIN, S. B. 2002. The fundamental role of epigenetic events in cancer. *Nat Rev Genet*, 3, 415-28.
- JONES, P. A. & LAIRD, P. W. 1999. Cancer epigenetics comes of age. *Nat Genet*, 21, 163-7.
- JOSEPH, R., ORLOV, Y. L., HUSS, M., SUN, W., KONG, S. L., UKIL, L., PAN, Y. F., LI, G., LIM, M., THOMSEN, J. S., RUAN, Y., CLARKE, N. D., PRABHAKAR, S., CHEUNG, E. & LIU, E. T. 2010. Integrative model of genomic factors for determining binding site selection by estrogen receptor-alpha. *Mol Syst Biol*, 6, 456.
- JOTHI, R., CUDDAPAH, S., BARSKI, A., CUI, K. & ZHAO, K. 2008. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res*, 36, 5221-31.
- KALA KARUN, A. & SUBU, S. 2013. BigTable, Dynamo & Cassandra – A Review. *International Journal of Electronics and Computer Science Engineering*, 2, 133-140.
- KALKMAN, S., BARENTSZ, M. W. & VAN DIEST, P. J. 2014. The effects of under 6 hours of formalin fixation on hormone receptor and HER2 expression in invasive breast cancer: a systematic review. *Am J Clin Pathol*, 142, 16-22.
- KANAVOS, P. 2006. The rising burden of cancer in the developing world. *Ann Oncol*, 17 Suppl 8, viii15-viii23.
- KAPPELMANN, M., BOSSERHOFF, A. & KUPHAL, S. 2014. AP-1/c-Jun transcription factors: regulation and function in malignant melanoma. *Eur J Cell Biol*, 93, 76-81.
- KAROLCHIK, D., HINRICHS, A. S., FUREY, T. S., ROSKIN, K. M., SUGNET, C. W., HAUSSLER, D. & KENT, W. J. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*, 32, D493-6.

- KASOWSKI, M., GRUBERT, F., HEFFELFINGER, C., HARIHARAN, M., ASABERE, A., WASZAK, S. M., HABEGGER, L., ROZOWSKY, J., SHI, M., URBAN, A. E., HONG, M. Y., KARCZEWSKI, K. J., HUBER, W., WEISSMAN, S. M., GERSTEIN, M. B., KORBEL, J. O. & SNYDER, M. 2010. Variation in transcription factor binding among humans. *Science*, 328, 232-5.
- KATAYAMA, T., SAISHO, K. & FUKUDA, A. Prototype of the device driver generation system for unix-like operating systems. *Principles of Software Evolution*, 2000. Proceedings. International Symposium on, 2000. IEEE, 302-310.
- KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M. & HAUSSLER, D. 2002a. The human genome browser at UCSC. *Genome Research*, 12, 996-1006.
- KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M. & HAUSSLER, D. 2002b. The human genome browser at UCSC. *Genome Res*, 12, 996-1006.
- KERPEDIJEV, P., FRELLSEN, J., LINDGREEN, S. & KROGH, A. 2014. Adaptable probabilistic mapping of short reads using position specific scoring matrices. *BMC Bioinformatics*, 15, 100.
- KHARCHENKO, P. V., TOLSTORUKOV, M. Y. & PARK, P. J. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol*, 26, 1351-9.
- KHUSHI, M. 2015. Benchmarking database performance for genomic data. *J Cell Biochem*, 116, 877-83.
- KHUSHI, M., CARPENTER, J. E., BALLEINE, R. L. & CLARKE, C. L. 2012a. Development of a data entry auditing protocol and quality assurance for a tissue bank database. *Cell Tissue Bank*, 13, 9-13.

- KHUSHI, M., CARPENTER, J. E., BALLEINE, R. L. & CLARKE, C. L. 2012b. Electronic biorepository application system: web-based software to manage receipt, peer review, and approval of researcher applications to a biobank. *Biopreserv Biobank*, 10, 37-44.
- KHUSHI, M., LIDDLE, C., CLARKE, C. L. & GRAHAM, J. D. 2014. Binding sites analyser (BiSA): software for genomic binding sites archiving and overlap analysis. *PLoS One*, 9, e87301.
- KILGORE, R. A. Open source initiatives for simulation software: multi-language, open-source modeling using the microsoft. NET architecture. Proceedings of the 34th conference on Winter simulation: exploring new frontiers, 2002. Winter Simulation Conference, 629-633.
- KIM, H., KIM, J., SELBY, H., GAO, D., TONG, T., PHANG, T. L. & TAN, A. C. 2011. A short survey of computational analysis methods in analysing ChIP-seq data. *Hum Genomics*, 5, 117-23.
- KIM, J. J., KURITA, T. & BULUN, S. E. 2013a. Progesterone action in endometrial cancer, endometriosis, uterine fibroids, and breast cancer. *Endocr Rev*, 34, 130-62.
- KIM, W., KIM, M. & JHO, E. H. 2013b. Wnt/beta-catenin signalling: from plasma membrane to nucleus. *Biochem J*, 450, 9-21.
- KING, D. C., TAYLOR, J., ELNITSKI, L., CHIAROMONTE, F., MILLER, W. & HARDISON, R. C. 2005. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res*, 15, 1051-60.
- KINSER, J. 2010. *Python for bioinformatics*, Jones & Bartlett Publishers.
- KITTLER, R., ZHOU, J., HUA, S., MA, L., LIU, Y., PENDLETON, E., CHENG, C., GERSTEIN, M. & WHITE, K. P. 2013. A comprehensive nuclear receptor network for breast cancer cells. *Cell Rep*, 3, 538-51.

- KLEIN, C. & VASSILEV, L. T. 2004. Targeting the p53-MDM2 interaction to treat cancer. *Br J Cancer*, 91, 1415-9.
- KRYSTKOWIAK, I., LENART, J., DEBSKI, K., KUTERBA, P., PETAS, M., KAMINSKA, B. & DABROWSKI, M. 2013. Nencki Genomics Database--Ensembl funcgen enhanced with intersections, user data and genome-wide TFBS motifs. *Database (Oxford)*, 2013, bat069.
- KULAKOVSKIY, I. V., BOEVA, V. A., FAVOROV, A. V. & MAKEEV, V. J. 2010. Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics*, 26, 2622-3.
- KUROSE, J. F. & ROSS, K. W. 2012. *Computer networking*, Pearson Education.
- KUTTIPPURATHU, L., HSING, M., LIU, Y., SCHMIDT, B., MASKELL, D. L., LEE, K., HE, A., PU, W. T. & KONG, S. W. 2011. CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments. *Bioinformatics*, 27, 715-7.
- KYRCHANOVA, O. & GEORGIEV, P. 2014. Chromatin insulators and long-distance interactions in Drosophila. *FEBS Lett*, 588, 8-14.
- LAAJALA, T. D., RAGHAV, S., TUOMELA, S., LAHESMAA, R., AITTOKALLIO, T. & ELO, L. L. 2009. A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC Genomics*, 10, 618.
- LACKIE, J. 2010. *A dictionary of biomedicine*, Oxford University Press.
- LADUNGA, I. 2010. An overview of the computational analyses and discovery of transcription factor binding sites. *Methods Mol Biol*, 674, 1-22.
- LAL, G., CONTRERAS, P. G., KULAK, M., WOODFIELD, G., BAIR, T., DOMANN, F. E. & WEIGEL, R. J. 2013. Human Melanoma cells over-express extracellular matrix 1 (ECM1) which is regulated by TFAP2C. *PLoS One*, 8, e73953.

- LAM, E. W., BROSENS, J. J., GOMES, A. R. & KOO, C. Y. 2013. Forkhead box proteins: tuning forks for transcriptional harmony. *Nat Rev Cancer*, 13, 482-95.
- LANDT, S. G., MARINOV, G. K., KUNDAJE, A., KHERADPOUR, P., PAULI, F., BATZOGLOU, S., BERNSTEIN, B. E., BICKEL, P., BROWN, J. B., CAYTING, P., CHEN, Y., DESALVO, G., EPSTEIN, C., FISHER-AYLOR, K. I., EUSKIRCHEN, G., GERSTEIN, M., GERTZ, J., HARTEMINK, A. J., HOFFMAN, M. M., IYER, V. R., JUNG, Y. L., KARMAKAR, S., KELLIS, M., KHARCHENKO, P. V., LI, Q., LIU, T., LIU, X. S., MA, L., MILOSAVLJEVIC, A., MYERS, R. M., PARK, P. J., PAZIN, M. J., PERRY, M. D., RAHA, D., REDDY, T. E., ROZOWSKY, J., SHORESH, N., SIDOW, A., SLATTERY, M., STAMATOYANNOPOULOS, J. A., TOLSTORUKOV, M. Y., WHITE, K. P., XI, S., FARNHAM, P. J., LIEB, J. D., WOLD, B. J. & SNYDER, M. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res*, 22, 1813-31.
- LANGMEAD, B. & SALZBERG, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9, 357-9.
- LANGMEAD, B., TRAPNELL, C., POP, M. & SALZBERG, S. L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10, R25.
- LATCHMAN, D. S. 2008. *Eukaryotic Transcription Factors*, New York, Academic Press.
- LAURELL, T., VANDERMEER, J. E., WENGER, A. M., GRIGELIONIENE, G., NORDENSKJOLD, A., ARNER, M., EKBLUM, A. G., BEJERANO, G., AHITUV, N. & NORDGREN, A. 2012. A novel 13 base pair insertion in the sonic hedgehog ZRS limb enhancer (ZRS/LMBR1) causes preaxial polydactyly with triphalangeal thumb. *Hum Mutat*, 33, 1063-6.

- LAWRENCE, J., CAMERON, D. & ARGYLE, D. 2015. Species differences in tumour responses to cancer chemotherapy. *Philos Trans R Soc Lond B Biol Sci*, 370.
- LAYER, R. M., SKADRON, K., ROBINS, G., HALL, I. M. & QUINLAN, A. R. 2013. Binary Interval Search: a scalable algorithm for counting interval intersections. *Bioinformatics*, 29, 1-7.
- LE GUEZENNEC, X., BRINKMAN, A. B., VERMEULEN, M., DENISSOV, S. G., GAZZIOLA, C., LOHRUM, M. E. & STUNNENBERG, H. G. 2005. Targeted discovery tools: proteomics and chromatin immunoprecipitation-on-chip. *BJU Int*, 96 Suppl 2, 16-22.
- LEAVITT, N. 2010. Will NoSQL Databases Live Up to Their Promise? *Computer*, 43, 12-14.
- LEBLANC, B. & MOSS, T. 2001. DNase I Footprinting. In: MOSS, T. (ed.) *DNA-Protein Interactions*. Humana Press.
- LEE, B. K. & IYER, V. R. 2012. Genome-wide studies of CCCTC-binding factor (CTCF) and cohesin provide insight into chromatin structure and regulation. *J Biol Chem*, 287, 30906-13.
- LEE, C. S., FRIEDMAN, J. R., FULMER, J. T. & KAESTNER, K. H. 2005. The initiation of liver development is dependent on Foxa transcription factors. *Nature*, 435, 944-7.
- LEE, T. I., JOHNSTONE, S. E. & YOUNG, R. A. 2006. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nat Protoc*, 1, 729-48.
- LEVINE, M., CATTOGLIO, C. & TJIAN, R. 2014. Looping back to leap forward: transcription enters a new era. *Cell*, 157, 13-25.
- LEVINE, M. & TJIAN, R. 2003. Transcription regulation and animal diversity. *Nature*, 424, 147-51.

- LI, H. 2011. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, 27, 718-9.
- LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25, 1754-60.
- LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & GENOME PROJECT DATA PROCESSING, S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078-9.
- LIM, B., PARK, J. L., KIM, H. J., PARK, Y. K., KIM, J. H., SOHN, H. A., NOH, S. M., SONG, K. S., KIM, W. H., KIM, Y. S. & KIM, S. Y. 2014. Integrative genomics analysis reveals the multilevel dysregulation and oncogenic characteristics of TEAD4 in gastric cancer. *Carcinogenesis*, 35, 1020-7.
- LIN, C. Y., VEGA, V. B., THOMSEN, J. S., ZHANG, T., KONG, S. L., XIE, M., CHIU, K. P., LIPOVICH, L., BARNETT, D. H., STOSI, F., YEO, A., GEORGE, J., KUZNETSOV, V. A., LEE, Y. K., CHARN, T. H., PALANISAMY, N., MILLER, L. D., CHEUNG, E., KATZENELLENBOGEN, B. S., RUAN, Y., BOURQUE, G., WEI, C. L. & LIU, E. T. 2007. Whole-genome cartography of estrogen receptor alpha binding sites. *PLoS Genet*, 3, e87.
- LIU, H. W., ZHANG, J., HEINE, G. F., ARORA, M., GULCIN OZER, H., ONTI-SRINIVASAN, R., HUANG, K. & PARVIN, J. D. 2012. Chromatin modification by SUMO-1 stimulates the promoters of translation machinery genes. *Nucleic Acids Res*, 40, 10172-86.
- LIU, T., ORTIZ, J., TAING, L., MEYER, C., LEE, B., ZHANG, Y., SHIN, H., WONG, S., MA, J., LEI, Y., PAPE, U., POIDINGER, M., CHEN, Y., YEUNG, K., BROWN, M., TURPAZ, Y. & LIU, X. S. 2011a. Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biology*, 12, R83.

- LIU, T., ORTIZ, J. A., TAING, L., MEYER, C. A., LEE, B., ZHANG, Y., SHIN, H., WONG, S. S., MA, J., LEI, Y., PAPE, U. J., POIDINGER, M., CHEN, Y., YEUNG, K., BROWN, M., TURPAZ, Y. & LIU, X. S. 2011b. Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol*, 12, R83.
- LOSADA, A. 2014. Cohesin in cancer: chromosome segregation and beyond. *Nat Rev Cancer*, 14, 389-93.
- LUNTER, G. & GOODSON, M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res*, 21, 936-9.
- LUPIEN, M., EECKHOUTE, J., MEYER, C. A., WANG, Q., ZHANG, Y., LI, W., CARROLL, J. S., LIU, X. S. & BROWN, M. 2008. FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell*, 132, 958-70.
- MACHANICK, P. & BAILEY, T. L. 2011. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, 27, 1696-7.
- MAHONY, S. & BENOS, P. V. 2007. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res*, 35, W253-8.
- MARIN, F. & DRAGOS, C. 2013. NoSQL and SQL Databases for Mobile Applications. Case Study: MongoDB versus PostgreSQL. *Informatica Economica Journal*, 17, 41-58.
- MASSA, H. A. & RIGGS, S. 2013. PostgreSQL in the database landscape.
- MASTON, G. A., EVANS, S. K. & GREEN, M. R. 2006. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*, 7, 29-59.
- MASTON, G. A., LANDT, S. G., SNYDER, M. & GREEN, M. R. 2012. Characterization of enhancer function from genome-wide analyses. *Annu Rev Genomics Hum Genet*, 13, 29-57.
- MCCARTY, B. 2004. *Learning Red Hat Enterprise Linux and Fedora*, O'Reilly Media, Inc.

- MCLEAN, C. Y., BRISTOR, D., HILLER, M., CLARKE, S. L., SCHAAR, B. T., LOWE, C. B., WENGER, A. M. & BEJERANO, G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol*, 28, 495-501.
- MEHTA, G. D., KUMAR, R., SRIVASTAVA, S. & GHOSH, S. K. 2013. Cohesin: functions beyond sister chromatid cohesion. *FEBS Lett*, 587, 2299-312.
- MEIJSING, S. H., PUFALL, M. A., SO, A. Y., BATES, D. L., CHEN, L. & YAMAMOTO, K. R. 2009. DNA binding site sequence directs glucocorticoid receptor structure and activity. *Science*, 324, 407-10.
- MENG, G. & VINGRON, M. 2014. Condition-specific target prediction from motifs and expression. *Bioinformatics*, 30, 1643-50.
- MESROUZE, Y., HAU, J. C., ERDMANN, D., ZIMMERMANN, C., FONTANA, P., SCHMELZLE, T. & CHENE, P. 2014. The surprising features of the TEAD4-Vgll1 protein-protein interaction. *Chembiochem*, 15, 537-42.
- MEYER, C. A., TANG, Q. & LIU, X. S. 2012. Minireview: applications of next-generation sequencing on studies of nuclear receptor regulation and function. *Mol Endocrinol*, 26, 1651-9.
- MEYERS, S. & LEE, M. 2011. *Learn OS X Lion*, Apress.
- MICROSOFT. 2013a. *Microsoft DreamSpark* [Online]. Available: <https://www.dreamspark.com/> [Accessed 06/11/2013].
- MICROSOFT. 2013b. *Microsoft SQL Server 2008 Management Studio Express* [Online]. Available: <http://www.microsoft.com/download/en/details.aspx?id=7593> [Accessed 06/11/2013].
- MICROSOFT. 2013c. *Microsoft WebsiteSpark* [Online]. Available: <http://www.microsoft.com/web/websitespark/> [Accessed 06/11/2013].
- MINGW. 2013. *MinGW | Minimalist GNU for Windows* [Online]. Available: <http://www.mingw.org/> [Accessed 13/07/2013 2013].

- MODEL, M. L. 2010. *Bioinformatics programming using python*, O'Reilly Media, Inc.
- MORTAZAVI, A., WILLIAMS, B. A., MCCUE, K., SCHAEFFER, L. & WOLD, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, 5, 621-8.
- MOTALLEBPOUR, M., AMEUR, A., REDDY BYSANI, M. S., PATRA, K., WALLERMAN, O., MANGION, J., BARKER, M., MCKERNAN, K., KOMOROWSKI, J. & WADELIUS, C. 2009a. Differential binding and co-binding pattern of FOXA1 and FOXA3 and their relation to H3K4me3 in HepG2 cells revealed by ChIP-seq. *Genome Biology*, 10, R129.
- MOTALLEBPOUR, M., AMEUR, A., REDDY BYSANI, M. S., PATRA, K., WALLERMAN, O., MANGION, J., BARKER, M. A., MCKERNAN, K. J., KOMOROWSKI, J. & WADELIUS, C. 2009b. Differential binding and co-binding pattern of FOXA1 and FOXA3 and their relation to H3K4me3 in HepG2 cells revealed by ChIP-seq. *Genome Biol*, 10, R129.
- MUKHOPADHYAY, A., DEPLANCKE, B., WALHOUT, A. J. & TISSENBAUM, H. A. 2008. Chromatin immunoprecipitation (ChIP) coupled to detection by quantitative real-time PCR to study transcription factor binding to DNA in *Caenorhabditis elegans*. *Nat Protoc*, 3, 698-709.
- MURGUIA, J. R. & SERRANO, R. 2012. New functions of protein kinase Gcn2 in yeast and mammals. *IUBMB Life*, 64, 971-4.
- NAKATSU, F., HASE, K. & OHNO, H. 2014. The Role of the Clathrin Adaptor AP-1: Polarized Sorting and Beyond. *Membranes (Basel)*, 4, 747-63.
- NAKSHATRI, H. & BADVE, S. 2009. FOXA1 in breast cancer. *Expert Rev Mol Med*, 11, e8.
- NEPH, S., KUEHN, M. S., REYNOLDS, A. P., HAUGEN, E., THURMAN, R. E., JOHNSON, A. K., RYNES, E., MAURANO, M. T., VIERSTRA, J., THOMAS,

- S., SANDSTROM, R., HUMBERT, R. & STAMATOYANNOPOULOS, J. A. 2012. BEDOPS: high-performance genomic feature operations. *Bioinformatics*, 28, 1919-20.
- NISHIMURA, H. & TIMOSSI, C. 2006. mono for cross-platform control system environment.
- NOVAK, P., NEUMANN, P., PECH, J., STEINHAISSL, J. & MACAS, J. 2013. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, 29, 792-3.
- O'NEILL, L. P. & TURNER, B. M. 2003. Immunoprecipitation of native chromatin: NChIP. *Methods*, 31, 76-82.
- OBIORAH, I. E., FAN, P., SENGUPTA, S. & JORDAN, V. C. 2014. Selective estrogen-induced apoptosis in breast cancer. *Steroids*, 90, 60-70.
- OGBOURNE, S. & ANTALIS, T. M. 1998. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochem J*, 331 (Pt 1), 1-14.
- OINN, T., ADDIS, M., FERRIS, J., MARVIN, D., SENGER, M., GREENWOOD, M., CARVER, T., GLOVER, K., POCOCK, M. R., WIPAT, A. & LI, P. 2004. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20, 3045-54.
- OKEGAWA, T., USHIO, K., IMAI, M., MORIMOTO, M. & HARA, T. 2013. Orphan nuclear receptor HNF4G promotes bladder cancer growth and invasion through the regulation of the hyaluronan synthase 2 gene. *Oncogenesis*, 2, e58.
- OLIPHANT, T. E. 2007. Python for scientific computing. *Computing in Science & Engineering*, 9, 10-20.
- ONG, C. T. & CORCES, V. G. 2012. Enhancers: emerging roles in cell fate specification. *EMBO Rep*, 13, 423-30.

- ORLANDO, V. 2000. Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem Sci*, 25, 99-104.
- ORLANDO, V., STRUTT, H. & PARO, R. 1997. Analysis of chromatin structure by in vivo formaldehyde cross-linking. *Methods*, 11, 205-14.
- ORR, N., LEMNRAU, A., COOKE, R., FLETCHER, O., TOMCZYK, K., JONES, M., JOHNSON, N., LORD, C. J., MITSOPOULOS, C., ZVELEBIL, M., MCDADE, S. S., BUCK, G., BLANCHER, C., CONSORTIUM, K. C., TRAINER, A. H., JAMES, P. A., BOJESEN, S. E., BOKMAND, S., NEVANLINNA, H., MATTSON, J., FRIEDMAN, E., LAITMAN, Y., PALLI, D., MASALA, G., ZANNA, I., OTTINI, L., GIANNINI, G., HOLLESTELLE, A., OUWELAND, A. M., NOVAKOVIC, S., KRAJC, M., GAGO-DOMINGUEZ, M., CASTELAO, J. E., OLSSON, H., HEDENFALK, I., EASTON, D. F., PHAROAH, P. D., DUNNING, A. M., BISHOP, D. T., NEUHAUSEN, S. L., STEELE, L., HOULSTON, R. S., GARCIA-CLOSAS, M., ASHWORTH, A. & SWERDLOW, A. J. 2012. Genome-wide association study identifies a common variant in RAD51B associated with male breast cancer risk. *Nat Genet*, 44, 1182-4.
- OUSTERHOUT, J. K. 1998. Scripting: Higher level programming for the 21st century. *Computer*, 31, 23-+.
- OVASKA, K., LYLTY, L., SAHU, B., JANNE, O. A. & HAUTANIEMI, S. 2013. Genomic region operation kit for flexible processing of deep sequencing data. *IEEE/ACM Trans Comput Biol Bioinform*, 10, 200-6.
- PABO, C. O. & SAUER, R. T. 1992. Transcription factors: structural families and principles of DNA recognition. *Annu Rev Biochem*, 61, 1053-95.
- PADILLA, A. & HAWKINS, T. 2011. Database Optimization. *Pro PHP Application Performance*. Springer.

- PAILA, U., CHAPMAN, B. A., KIRCHNER, R. & QUINLAN, A. R. 2013. GEMINI: integrative exploration of genetic variation and genome annotations. *PLoS Comput Biol*, 9, e1003153.
- PALSTRA, R. J., DE LAAT, W. & GROSVELD, F. 2008. Beta-globin regulation and long-range interactions. *Adv Genet*, 61, 107-42.
- PALSTRA, R. J. & GROSVELD, F. 2012. Transcription factor binding at enhancers: shaping a genomic regulatory landscape in flux. *Front Genet*, 3, 195.
- PANASCI, L., ALOYZ, R. & ALAOUI-JAMALI, M. 2012. Advances in DNA Repair in Cancer Therapy. 1 ed. Dordrecht: Springer.
- PARK, P. J. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10, 669-80.
- PARTANEN, A., MOTOYAMA, J. & HUI, C. C. 1999. Developmentally regulated expression of the transcriptional cofactors/histone acetyltransferases CBP and p300 during mouse embryogenesis. *Int J Dev Biol*, 43, 487-94.
- PAULSON, L. D. 2004. Open source databases move into the marketplace. *Computer*, 37, 13-15.
- PEARSON, W. R. 1994. Using the FASTA program to search protein and DNA sequence databases. *Computer Analysis of Sequence Data*. Springer.
- PEKOWSKA, A., BENOUKRAF, T., ZACARIAS-CABEZA, J., BELHOCINE, M., KOCH, F., HOLOTA, H., IMBERT, J., ANDRAU, J. C., FERRIER, P. & SPICUGLIA, S. 2011. H3K4 tri-methylation provides an epigenetic signature of active enhancers. *EMBO J*, 30, 4198-210.
- PENAULT-LLORCA, F. & VIALE, G. 2012. Pathological and molecular diagnosis of triple-negative breast cancer: a clinical perspective. *Ann Oncol*, 23 Suppl 6, vi19-22.

- PEPKE, S., WOLD, B. & MORTAZAVI, A. 2009. Computation for ChIP-seq and RNA-seq studies. *Nat Methods*, 6, S22-32.
- PHILLIPS, T. 2008. Small non-coding RNA and gene expression. *Nature Education*, 1.
- PHILLIPS, T. & SHAW, K. 2008. Chromatin remodeling in eukaryotes. *Nature Education*, 1.
- PRLIĆ, A., YATES, A., BLIVEN, S. E., ROSE, P. W., JACOBSEN, J., TROSHIN, P. V., CHAPMAN, M., GAO, J., KOH, C. H. & FOISY, S. 2012. BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics*, 28, 2693-2695.
- PTASHNE, M. & GANN, A. 1997. Transcriptional activation by recruitment. *Nature*, 386, 569-77.
- QUINLAN, A. R. & HALL, I. M. 2010a. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26, 841-2.
- QUINLAN, A. R. & HALL, I. M. 2010b. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26, 841-842.
- RAHA, D., WANG, Z., MOQTADERI, Z., WU, L., ZHONG, G., GERSTEIN, M., STRUHL, K. & SNYDER, M. 2010. Close association of RNA polymerase II and many transcription factors with Pol III genes. *Proc Natl Acad Sci U S A*, 107, 3639-44.
- RAM, O., GOREN, A., AMIT, I., SHORESH, N., YOSEF, N., ERNST, J., KELLIS, M., GYMREK, M., ISSNER, R., COYNE, M., DURHAM, T., ZHANG, X., DONAGHEY, J., EPSTEIN, C. B., REGEV, A. & BERNSTEIN, B. E. 2011. Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. *Cell*, 147, 1628-39.
- RAY, S., SIMION, B. & BROWN, A. D. Jackpine: A benchmark to evaluate spatial database performance. Data Engineering (ICDE), 2011 IEEE 27th International Conference on, 11-16 April 2011 2011. 1139-1150.

- REN, B., ROBERT, F., WYRICK, J. J., APARICIO, O., JENNINGS, E. G., SIMON, I., ZEITLINGER, J., SCHREIBER, J., HANNETT, N., KANIN, E., VOLKERT, T. L., WILSON, C. J., BELL, S. P. & YOUNG, R. A. 2000. Genome-wide location and function of DNA binding proteins. *Science*, 290, 2306-9.
- RENAUD, G., NEVES, P., FOLADOR, E. L., FERREIRA, C. G. & PASSETTI, F. 2011. Segtor: rapid annotation of genomic coordinates and single nucleotide variations using segment trees. *PLoS One*, 6, e26715.
- RIETHOVEN, J. J. 2010. Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. *Methods Mol Biol*, 674, 33-42.
- ROBERTSON, G., HIRST, M., BAINBRIDGE, M., BILENKY, M., ZHAO, Y., ZENG, T., EUSKIRCHEN, G., BERNIER, B., VARHOL, R., DELANEY, A., THIESSEN, N., GRIFFITH, O. L., HE, A., MARRA, M., SNYDER, M. & JONES, S. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*, 4, 651-7.
- ROH, T. Y., CUDDAPAH, S., CUI, K. & ZHAO, K. 2006. The genomic landscape of histone modifications in human T cells. *Proc Natl Acad Sci U S A*, 103, 15782-7.
- ROYCE, T. E., ROZOWSKY, J. S., BERTONE, P., SAMANTA, M., STOLC, V., WEISSMAN, S., SNYDER, M. & GERSTEIN, M. 2005. Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends Genet*, 21, 466-75.
- ROZOWSKY, J., EUSKIRCHEN, G., AUERBACH, R. K., ZHANG, Z. D., GIBSON, T., BJORNSON, R., CARRIERO, N., SNYDER, M. & GERSTEIN, M. B. 2009. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol*, 27, 66-75.

- RUBIO, E. D., REISS, D. J., WELCSH, P. L., DISTECHE, C. M., FILIPPOVA, G. N., BALIGA, N. S., AEBERSOLD, R., RANISH, J. A. & KRUMM, A. 2008. CTCF physically links cohesin to chromatin. *Proc Natl Acad Sci U S A*, 105, 8309-14.
- RUSS, B. E., DENTON, A. E., HATTON, L., CROOM, H., OLSON, M. R. & TURNER, S. J. 2012. Defining the molecular blueprint that drives CD8(+) T cell differentiation in response to infection. *Front Immunol*, 3, 371.
- SAHU, B., LAAKSO, M., OVASKA, K., MIRTTI, T., LUNDIN, J., RANNIKKO, A., SANKILA, A., TURUNEN, J. P., LUNDIN, M., KONSTI, J., VESTERINEN, T., NORDLING, S., KALLIONIEMI, O., HAUTANIEMI, S. & JANNE, O. A. 2011. Dual role of FoxA1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer. *EMBO J*, 30, 3962-76.
- SALEHNIA, M. & ZAVAREH, S. 2013. The effects of progesterone on oocyte maturation and embryo development. *Int J Fertil Steril*, 7, 74-81.
- SANDMANN, T., JENSEN, L. J., JAKOBSEN, J. S., KARZYNSKI, M. M., EICHENLAUB, M. P., BORK, P. & FURLONG, E. E. 2006. A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. *Dev Cell*, 10, 797-807.
- SCHATZ, M. C. 2010. The missing graphical user interface for genomics. *Genome Biol*, 11, 128.
- SCHMIDT, D., SCHWALIE, P. C., ROSS-INNES, C. S., HURTADO, A., BROWN, G. D., CARROLL, J. S., FLICEK, P. & ODOM, D. T. 2010. A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res*, 20, 578-88.
- SHABALINA, S. A. & KOONIN, E. V. 2008. Origins and evolution of eukaryotic RNA interference. *Trends Ecol Evol*, 23, 578-87.

- SHAO, R., CAO, S., WANG, X., FENG, Y. & BILLIG, H. 2014. The elusive and controversial roles of estrogen and progesterone receptors in human endometriosis. *Am J Transl Res*, 6, 104-13.
- SHARE, N. M. 2013. Market Share for Mobile and Desktop.
- SHELDON, R. & MOES, G. 2005. *Beginning MySQL*, Wiley.
- SHEN, L., CHOI, I., NESTLER, E. J. & WON, K. J. 2013. Human Transcriptome and Chromatin Modifications: An ENCODE Perspective. *Genomics Inform*, 11, 60-7.
- SHIKAMA, N., LUTZ, W., KRETZSCHMAR, R., SAUTER, N., ROTH, J. F., MARINO, S., WITTEWER, J., SCHEIDWEILER, A. & ECKNER, R. 2003. Essential function of p300 acetyltransferase activity in heart, lung and small intestine formation. *EMBO J*, 22, 5175-85.
- SHLYUEVA, D., STAMPFEL, G. & STARK, A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet*, 15, 272-86.
- SIM, S. E., EASTERBROOK, S. & HOLT, R. C. Using benchmarking to advance research: A challenge to software engineering. Proceedings of the 25th International Conference on Software Engineering, 2003. IEEE Computer Society, 74-83.
- SIMMEN, K. A., NEWELL, A., ROBINSON, M., MILLS, J. S., CANNING, G., HANDA, R., PARKES, K., BORKAKOTI, N. & JUPP, R. 1997. Protein interactions in the herpes simplex virus type 1 VP16-induced complex: VP16 peptide inhibition and mutational analysis of host cell factor requirements. *J Virol*, 71, 3886-94.
- SIMMONS, D. 2008. Epigenetic influence and disease. *Nature Education*, 1.
- SINGH, S., JALANDHAR, I. & DHIR, R. 2013. Simulation Based Performance Comparison of Proactive and Reactive Routing Protocols in MANET using FTP and HTTP Traffics.

- SITES, R. L., CHERNOFF, A., KIRK, M. B., MARKS, M. P. & ROBINSON, S. G. 1993. Binary Translation. *Communications of the Acm*, 36, 69-81.
- SMITH, E. & SHILATIFARD, A. 2014. Enhancer biology and enhanceropathies. *Nat Struct Mol Biol*, 21, 210-9.
- SOCCIO, R. E., TUTEJA, G., EVERETT, L. J., LI, Z., LAZAR, M. A. & KAESTNER, K. H. 2011. Species-specific strategies underlying conserved functions of metabolic transcription factors. *Mol Endocrinol*, 25, 694-706.
- SOLOMON, M. J., LARSEN, P. L. & VARSHAVSKY, A. 1988. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell*, 53, 937-47.
- SPILIANAKIS, C. G., LALIOTI, M. D., TOWN, T., LEE, G. R. & FLAVELL, R. A. 2005. Interchromosomal associations between alternatively expressed loci. *Nature*, 435, 637-45.
- SPITZ, F. & FURLONG, E. E. 2012. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet*, 13, 613-26.
- STAJICH, J. E., BLOCK, D., BOULEZ, K., BRENNER, S. E., CHERVITZ, S. A., DAGDIGIAN, C., FUELLEN, G., GILBERT, J. G., KORF, I., LAPP, H., LEHVASLAIHO, H., MATSALLA, C., MUNGALL, C. J., OSBORNE, B. I., POCOCK, M. R., SCHATTNER, P., SENGER, M., STEIN, L. D., STUPKA, E., WILKINSON, M. D. & BIRNEY, E. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12, 1611-8.
- STAJICH, J. E. & LAPP, H. 2006a. Open source tools and toolkits for bioinformatics: significance, and where are we? *Brief Bioinform*, 7, 287-96.
- STAJICH, J. E. & LAPP, H. 2006b. Open source tools and toolkits for bioinformatics: significance, and where are we? *BRIEFINGS IN BIOINFORMATICS*, 7, 287-296.

- STOWER, H. 2011. Gene regulation: Resolving transcription factor binding. *Nat Rev Genet*, 13, 71.
- STRÖM, A., HARTMAN, J., FOSTER, J. S., KIETZ, S., WIMALASENA, J. & GUSTAFSSON, J.-Å. 2004. Estrogen receptor β inhibits 17 β -estradiol-stimulated proliferation of the breast cancer cell line T47D. *Proceedings of the National Academy of Sciences of the United States of America*, 101, 1566-1571.
- SU, H., MA, X., HUANG, Y., HAN, H., ZOU, Y. & HUANG, W. 2015. JARID1B deletion induced apoptosis in Jeko-1 and HL-60 cell lines. *Int J Clin Exp Pathol*, 8, 171-83.
- TANG, Q., CHEN, Y., MEYER, C., GEISTLINGER, T., LUPIEN, M., WANG, Q., LIU, T., ZHANG, Y., BROWN, M. & LIU, X. S. 2011. A comprehensive view of nuclear receptor cancer cistromes. *Cancer Res*, 71, 6940-7.
- TANIGUCHI, M., FUJIWARA, K., NAKAI, Y., OZAKI, T., KOSHIKAWA, N., TOSHIO, K., KATABA, M., OGUNI, A., MATSUDA, H., YOSHIDA, Y., TOKUHASHI, Y., FUKUDA, N., UENO, T., SOMA, M. & NAGASE, H. 2014. Inhibition of malignant phenotypes of human osteosarcoma cells by a gene silencer, a pyrrole-imidazole polyamide, which targets an E-box motif. *FEBS Open Bio*, 4, 328-34.
- TANIGUCHI, Y. 2014. Hox transcription factors: modulators of cell-cell and cell-extracellular matrix adhesion. *Biomed Res Int*, 2014, 591374.
- TASLIM, C., WU, J., YAN, P., SINGER, G., PARVIN, J., HUANG, T., LIN, S. & HUANG, K. 2009. Comparative study on ChIP-seq data: normalization and binding pattern characterization. *Bioinformatics*, 25, 2334-40.
- TAYLOR, J., SCHENCK, I., BLANKENBERG, D. & NEKRUTENKO, A. 2007. Using Galaxy to Perform Large-Scale Interactive Data Analyses. *Current protocols in bioinformatics*, 10.5. 1-10.5. 25.

- THOMAS-CHOLLIER, M., HERRMANN, C., DEFRANCE, M., SAND, O., THIEFFRY, D. & VAN HELDEN, J. 2012. RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res*, 40, e31.
- TISDALL, J. 2009. *Beginning Perl for bioinformatics*, O'Reilly Media, Inc.
- TISDALL, J. 2010. *Mastering Perl for bioinformatics*, O'Reilly.
- TRISTAN, M., OROZCO, L. J., STEED, A., RAMIREZ-MORERA, A. & STONE, P. 2012. Mifepristone for uterine fibroids. *Cochrane Database Syst Rev*, 8, CD007687.
- TSAI, M. J. & O'MALLEY, B. W. 1994. Molecular mechanisms of action of steroid/thyroid receptor superfamily members. *Annu Rev Biochem*, 63, 451-86.
- TSAI, W. W., WANG, Z., YIU, T. T., AKDEMIR, K. C., XIA, W., WINTER, S., TSAI, C. Y., SHI, X., SCHWARZER, D., PLUNKETT, W., ARONOW, B., GOZANI, O., FISCHLE, W., HUNG, M. C., PATEL, D. J. & BARTON, M. C. 2010. TRIM24 links a non-canonical histone signature to breast cancer. *Nature*, 468, 927-32.
- TSANG, J. C., GAO, X., LU, L. & LIU, P. 2014. Cellular reprogramming by transcription factor engineering. *Curr Opin Genet Dev*, 28, 1-9.
- TSIRIGOS, A., HAIMINEN, N., BILAL, E. & UTRO, F. 2012. GenomicTools: a computational platform for developing high-throughput analytics in genomics. *Bioinformatics*, 28, 282-3.
- TUTEJA, G., WHITE, P., SCHUG, J. & KAESTNER, K. H. 2009. Extracting transcription factor targets from ChIP-Seq data. *Nucleic Acids Res*, 37, e113.
- VALOUEV, A., JOHNSON, D. S., SUNDQUIST, A., MEDINA, C., ANTON, E., BATZOGLOU, S., MYERS, R. M. & SIDOW, A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods*, 5, 829-34.

- VAN'T VEER, L. J., DAI, H., VAN DE VIJVER, M. J., HE, Y. D., HART, A. A., MAO, M., PETERSE, H. L., VAN DER KOOY, K., MARTON, M. J. & WITTEVEEN, A. T. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415, 530-536.
- VASSILEV, L. T., VU, B. T., GRAVES, B., CARVAJAL, D., PODLASKI, F., FILIPOVIC, Z., KONG, N., KAMMLOTT, U., LUKACS, C., KLEIN, C., FOTOUHI, N. & LIU, E. A. 2004. In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science*, 303, 844-8.
- VENEMA, V., MESTRE, O., AGUILAR, E., AUER, I., GUIJARRO, J., DOMONKOS, P., VERTACNIK, G., SZENTIMREY, T., STEPANEK, P. & ZAHRADNICEK, P. Benchmarking homogenization algorithms for monthly data. AIP Conference Proceedings, 2013. 1060.
- VINCKEVICIUS, A. & CHAKRAVARTI, D. 2012a. Chromatin immunoprecipitation: advancing analysis of nuclear hormone signaling. *Journal of Molecular Endocrinology*, 49, R113-R123.
- VINCKEVICIUS, A. & CHAKRAVARTI, D. 2012b. Chromatin immunoprecipitation: advancing analysis of nuclear hormone signaling. *J Mol Endocrinol*, 49, R113-23.
- WAGNER, E. F. 2010. Bone development and inflammatory disease is regulated by AP-1 (Fos/Jun). *Ann Rheum Dis*, 69 Suppl 1, i86-88.
- WALL, L., CHRISTIANSEN, T. & ORWANT, J. 2000. *Programming perl*, O'Reilly.
- WALLERMAN, O., MOTALLEBPOUR, M., ENROTH, S., PATRA, K., BYSANI, M. S., KOMOROWSKI, J. & WADELIUS, C. 2009. Molecular interactions between HNF4a, FOXA2 and GABP identified at regulatory DNA elements through ChIP-sequencing. *Nucleic Acids Res*, 37, 7498-508.
- WANG, C., MAYER, J. A., MAZUMDAR, A., FERTUCK, K., KIM, H., BROWN, M. & BROWN, P. H. 2011a. Estrogen induces c-myc gene expression via an upstream

- enhancer activated by the estrogen receptor and the AP-1 transcription factor. *Mol Endocrinol*, 25, 1527-38.
- WANG, C., ZHANG, M. Q. & ZHANG, Z. 2013. Computational identification of active enhancers in model organisms. *Genomics Proteomics Bioinformatics*, 11, 142-50.
- WANG, D., GARCIA-BASSETS, I., BENNER, C., LI, W., SU, X., ZHOU, Y., QIU, J., LIU, W., KAIKKONEN, M. U., OHGI, K. A., GLASS, C. K., ROSENFELD, M. G. & FU, X. D. 2011b. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature*, 474, 390-4.
- WANG, J., ZHUANG, J., IYER, S., LIN, X., WHITFIELD, T. W., GREVEN, M. C., PIERCE, B. G., DONG, X., KUNDAJE, A., CHENG, Y., RANDO, O. J., BIRNEY, E., MYERS, R. M., NOBLE, W. S., SNYDER, M. & WENG, Z. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res*, 22, 1798-812.
- WANG, L. & DI, L. J. 2014. BRCA1 and estrogen/estrogen receptor in breast cancer: where they interact? *Int J Biol Sci*, 10, 566-75.
- WANG, L., MAO, Y., DU, G., HE, C. & HAN, S. 2015. Overexpression of JARID1B is associated with poor prognosis and chemotherapy resistance in epithelial ovarian cancer. *Tumour Biol*, 36, 2465-72.
- WASSERMAN, W. W., PALUMBO, M., THOMPSON, W., FICKETT, J. W. & LAWRENCE, C. E. 2000. Human-mouse genome comparisons to locate regulatory sites. *Nat Genet*, 26, 225-8.
- WASSERMAN, W. W. & SANDELIN, A. 2004. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet*, 5, 276-87.
- WATANABE, H., FRANCIS, J. M., WOO, M. S., ETEMAD, B., LIN, W., FRIES, D. F., PENG, S., SNYDER, E. L., TATA, P. R., IZZO, F., SCHINZEL, A. C., CHO, J., HAMMERMAN, P. S., VERHAAK, R. G., HAHN, W. C., RAJAGOPAL, J.,

- JACKS, T. & MEYERSON, M. 2013. Integrated cisomic and expression analysis of amplified NKX2-1 in lung adenocarcinoma identifies LMO3 as a functional transcriptional target. *Genes Dev*, 27, 197-210.
- WEI, G., WEI, L., ZHU, J., ZANG, C., HU-LI, J., YAO, Z., CUI, K., KANNO, Y., ROH, T. Y., WATFORD, W. T., SCHONES, D. E., PENG, W., SUN, H. W., PAUL, W. E., O'SHEA, J. J. & ZHAO, K. 2009. Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells. *Immunity*, 30, 155-67.
- WELBORN, W. J., VAN DRIEL, M. A., JANSSEN-MEGENS, E. M., VAN HEERINGEN, S. J., SWEEP, F. C., SPAN, P. N. & STUNNENBERG, H. G. 2009. ChIP-Seq of ERalpha and RNA polymerase II defines genes differentially responding to ligands. *EMBO J*, 28, 1418-28.
- WELLING, L. & THOMSON, L. 2003. *PHP and MySQL Web development*, Sams Publishing.
- WHITINGTON, T., FRITH, M. C., JOHNSON, J. & BAILEY, T. L. 2011. Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res*, 39, e98.
- WILBANKS, E. G. & FACCIOTTI, M. T. 2010. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One*, 5, e11471.
- WINTERBOTTOM, P. & WILKINSON, T. MESHIX: a UNIX like operating system for distributed machines,". UKUUG Summer Conference Proceedings, 1990. Citeseer, 237-246.
- WOLF, E., LIN, C. Y., EILERS, M. & LEVENS, D. L. 2015. Taming of the beast: shaping Myc-dependent amplification. *Trends Cell Biol*, 25, 241-8.
- WOODFIELD, G. W., CHEN, Y., BAIR, T. B., DOMANN, F. E. & WEIGEL, R. J. 2010. Identification of primary gene targets of TFAP2C in hormone responsive breast carcinoma cells. *Genes Chromosomes Cancer*, 49, 948-62.

- WRAY, G. A. 2007. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet*, 8, 206-16.
- XIA, Y., CHANG, T., WANG, Y., LIU, Y., LI, W., LI, M. & FAN, H. Y. 2014. YAP promotes ovarian cancer cell tumorigenesis and is indicative of a poor prognosis for ovarian cancer patients. *PLoS One*, 9, e91770.
- XIANG, Y., ZHU, Z., HAN, G., YE, X., XU, B., PENG, Z., MA, Y., YU, Y., LIN, H., CHEN, A. P. & CHEN, C. D. 2007. JARID1B is a histone H3 lysine 4 demethylase up-regulated in prostate cancer. *Proc Natl Acad Sci U S A*, 104, 19226-31.
- XIE, W., SCHULTZ, M. D., LISTER, R., HOU, Z., RAJAGOPAL, N., RAY, P., WHITAKER, J. W., TIAN, S., HAWKINS, R. D., LEUNG, D., YANG, H., WANG, T., LEE, A. Y., SWANSON, S. A., ZHANG, J., ZHU, Y., KIM, A., NERY, J. R., URICH, M. A., KUAN, S., YEN, C. A., KLUGMAN, S., YU, P., SUKNUNTHA, K., PROPSON, N. E., CHEN, H., EDSALL, L. E., WAGNER, U., LI, Y., YE, Z., KULKARNI, A., XUAN, Z., CHUNG, W. Y., CHI, N. C., ANTOSIEWICZ-BOURGET, J. E., SLUKVIN, I., STEWART, R., ZHANG, M. Q., WANG, W., THOMSON, J. A., ECKER, J. R. & REN, B. 2013. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*, 153, 1134-48.
- XU, J. & GÜTING, R. H. 2012. GMOBench: A Benchmark for Generic Moving Objects. *Informatik-Report 362, Fernuniversität Hagen*.
- XU, W., CHEN, H., DU, K., ASAHARA, H., TINI, M., EMERSON, B. M., MONTMINY, M. & EVANS, R. M. 2001. A transcriptional switch mediated by cofactor methylation. *Science*, 294, 2507-11.
- YADAV, B. S., SHARMA, S. C., CHANANA, P. & JHAMB, S. 2014. Systemic treatment strategies for triple-negative breast cancer. *World J Clin Oncol*, 5, 125-33.

- YAMAMOTO, S., WU, Z., RUSSNES, H. G., TAKAGI, S., PELUFFO, G., VASKE, C., ZHAO, X., MOEN VOLLAN, H. K., MARUYAMA, R., EKRAM, M. B., SUN, H., KIM, J. H., CARVER, K., ZUCCA, M., FENG, J., ALMENDRO, V., BESSARABOVA, M., RUEDA, O. M., NIKOLSKY, Y., CALDAS, C., LIU, X. S. & POLYAK, K. 2014. JARID1B is a luminal lineage-driving oncogene in breast cancer. *Cancer Cell*, 25, 762-77.
- YAMANE, K., TATEISHI, K., KLOSE, R. J., FANG, J., FABRIZIO, L. A., ERDJUMENT-BROMAGE, H., TAYLOR-PAPADIMITRIOU, J., TEMPST, P. & ZHANG, Y. 2007. PLU-1 is an H3K4 demethylase involved in transcriptional repression and breast cancer cell proliferation. *Mol Cell*, 25, 801-12.
- YANG, C. S., CHANG, K. Y. & RANA, T. M. 2014. Genome-wide functional analysis reveals factors needed at the transition steps of induced reprogramming. *Cell Rep*, 8, 327-37.
- YANG, L., YANG, H., LI, J., HAO, J. & QIAN, J. 2013. ppENK Gene Methylation Status in the Development of Pancreatic Carcinoma. *Gastroenterol Res Pract*, 2013, 130927.
- YERUSHALMI, G. M., GILBOA, Y., JAKOBSON-SETTON, A., TADIR, Y., GOLDCHMIT, C., KATZ, D. & SEIDMAN, D. S. 2014. Vaginal mifepristone for the treatment of symptomatic uterine leiomyomata: an open-label study. *Fertil Steril*, 101, 496-500.
- YIN, P., ROQUEIRO, D., HUANG, L., OWEN, J. K., XIE, A., NAVARRO, A., MONSIVAIS, D., COON, J. S. T., KIM, J. J., DAI, Y. & BULUN, S. E. 2012. Genome-wide progesterone receptor binding: cell type-specific and shared mechanisms in T47D breast cancer cells and primary leiomyoma cells. *PLoS One*, 7, e29021.

- YU, J., YU, J., MANI, R. S., CAO, Q., BRENNER, C. J., CAO, X., WANG, X., WU, L., LI, J., HU, M., GONG, Y., CHENG, H., LAXMAN, B., VELLAICHAMY, A., SHANKAR, S., LI, Y., DHANASEKARAN, S. M., MOREY, R., BARRETTE, T., LONIGRO, R. J., TOMLINS, S. A., VARAMBALLY, S., QIN, Z. S. & CHINNAIYAN, A. M. 2010. An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell*, 17, 443-54.
- YUSUFZAI, T. M., TAGAMI, H., NAKATANI, Y. & FELSENFELD, G. 2004. CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Mol Cell*, 13, 291-8.
- ZAMMATARO, L., DEMOLFETTA, R., BUCCI, G., CEOL, A. & MULLER, H. 2014. AnnotateGenomicRegions: a web application. *BMC Bioinformatics*, 15 Suppl 1, S8.
- ZANG, C., SCHONES, D. E., ZENG, C., CUI, K., ZHAO, K. & PENG, W. 2009. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, 25, 1952-8.
- ZENTNER, G. E., TESAR, P. J. & SCACHERI, P. C. 2011. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res*, 21, 1273-83.
- ZENZ, R., EFERL, R., SCHEINECKER, C., REDLICH, K., SMOLEN, J., SCHONTHALER, H. B., KENNER, L., TSCHACHLER, E. & WAGNER, E. F. 2008. Activator protein 1 (Fos/Jun) functions in inflammatory bone and skin disease. *Arthritis Res Ther*, 10, 201.
- ZHANG, X., OUYANG, S., KONG, X., LIANG, Z., LU, J., ZHU, K., ZHAO, D., ZHENG, M., JIANG, H., LIU, X., MARMORSTEIN, R. & LUO, C. 2014.

Catalytic mechanism of histone acetyltransferase p300: from the proton transfer to acetylation reaction. *J Phys Chem B*, 118, 2009-19.

ZHANG, Y., LIU, T., MEYER, C. A., EECKHOUTE, J., JOHNSON, D. S., BERNSTEIN, B. E., NUSBAUM, C., MYERS, R. M., BROWN, M., LI, W. & LIU, X. S. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, 9, R137.

ZHANG, Z., CHANG, C. W., GOH, W. L., SUNG, W. K. & CHEUNG, E. 2011.

CENTDIST: discovery of co-associated factors by motif distribution. *Nucleic Acids Res*, 39, W391-9.

Appendix: Publications

Benchmarking Database Performance for Genomic Data

Matloob Khushi^{1,2*}

¹Bioinformatics Unit, Children's Medical Research Institute, Westmead, NSW, Australia

²Centre for Cancer Research, Westmead Millennium Institute; Sydney Medical School, Westmead, University of Sydney, Sydney, Australia

ABSTRACT

Genomic regions represent features such as gene annotations, transcription factor binding sites and epigenetic modifications. Performing various genomic operations such as identifying overlapping/non-overlapping regions or nearest gene annotations are common research needs. The data can be saved in a database system for easy management, however, there is no comprehensive database built-in algorithm at present to identify overlapping regions. Therefore I have developed a novel region-mapping (RegMap) SQL-based algorithm to perform genomic operations and have benchmarked the performance of different databases. Benchmarking identified that PostgreSQL extracts overlapping regions much faster than MySQL. Insertion and data uploads in PostgreSQL were also better, although general searching capability of both databases was almost equivalent. In addition, using the algorithm pair-wise, overlaps of > 1000 datasets of transcription factor binding sites and histone marks, collected from previous publications, were reported and it was found that HNF4G significantly co-locates with cohesin subunit STAG1 (SA1). *J. Cell. Biochem.* 116: 877–883, 2015. © 2015 Wiley Periodicals, Inc.

KEY WORDS: TRANSCRIPTION FACTOR BINDING SITES; EPIGENETIC MODIFICATIONS; DATABASE BENCHMARKING; MANAGING GENOMIC LOCATIONS DATA; REGMAP

The recent revolution in whole genome census approaches has seen an exponential increase in available data sets describing genomic features, such as transcription factor binding sites and histone modifications. Curation of such data and identifying relationships, such as overlaps in genomic features and closest gene annotation, are fundamental tasks in this research [Meyer et al., 2012]. The files containing such genomic data usually have chromosomal location (chromosome, start and end) information in them. A number of tools such as Galaxy [Goecks et al., 2010], BedTools, GenomicTools [Tsirigos et al., 2012] and BEDOPS Tools [Neph et al., 2012] have been developed to find overlapping/non-overlapping nearby regions [Quinlan and Hall, 2010; Neph et al., 2012; Zammataro et al., 2014]. The relationships among the files are usually manually managed. With exponential growth in available genomic information, managing manual relationship and curation of such files are becoming more cumbersome day by day. These relationships and curation can be better managed using a relational database such as Microsoft SQL Server, Oracle, MySQL or PostgreSQL; however, there is no dedicated published algorithm available that is natively built into a database system to operate on the genomic features. I have therefore developed a novel algorithm Region Mapping (RegMap) that operates on genomic locations natively in the database and have benchmarked the performance of

two major open-source free databases PostgreSQL and MySQL. I have also compared the RegMap performance against database built-in spatial functions which provide very limited functionality.

METHODS

Two genomic regions (genomic intervals) are said to intersect or overlap if both intervals share at least one base pair in common on a chromosome. Chromosomes were saved as character data-type and start and end coordinates as integer data-type for the RegMap algorithm. To compare performance with database built-in spatial functions the coordinates were saved as linear spatial data-type. Genomic regions were saved in the *Regions* table and were linked with the *RegionDesc* table where annotation of the regions was saved, thus simulating a production usage. Each region in the *Regions* table was automatically assigned a unique database id (Primary Key). The start coordinates of the genomic regions were indexed from 0, according to UCSC recommendations (<http://genome.ucsc.edu/>) to speed up calculations, therefore region length was calculated by subtracting the start from the end coordinate.

RegMap generates all the required objects in a working database. The algorithm was developed in native SQL (Structured Query Language) and is therefore compatible with all SQL databases.

*Correspondence to: Matloob Khushi, Bioinformatics Unit, Children's Medical Research Institute, 214 Hawkesbury Road, Westmead, NSW 2145, Australia. Email: mkhushi@uni.sydney.edu.au

Manuscript Received: 17 November 2014; Manuscript Accepted: 16 December 2014

Accepted manuscript online in Wiley Online Library (wileyonlinelibrary.com): 5 January 2015

DOI 10.1002/jcb.25049 • © 2015 Wiley Periodicals, Inc.

A total of 1005 datasets of transcription factor binding sites and histone marks from previous publications on human and mouse assemblies were collected, including hg 19, hg 18, mm9 or mm8. This 'Knowledge Base' was used to perform search benchmarking.

All testing and benchmarking were performed on PostgreSQL 9.0 and MySQL Community Server 5.6.15 GPL (x86_64) installed on a personal computer of 4 core 2.4 GHz processor with 8 GB memory. MySQL Server supports a number of storage engines, however I have benchmarked performance for two widely used InnoDB and MyISAM storage engines [Sheldon and Moes, 2005; Padilla and Hawkins, 2011]. The results of 100 simulations were averaged for all operations. RegMap code was run in MySQL Workbench 6.1 for MySQL server and in pgAdmin III 1.81 for PostgreSQL benchmarking maintaining the default settings of each database. The default random region size was set to 500, however, this setting can be changed in the script.

RESULTS

DEVELOPMENT OF THE RegMap ALGORITHM

RegMap finds overlapping or non-overlapping regions by calculating the number of bases common or away between two regions. Therefore, a variable 'bp overlap' was devised which was calculated positive (shown as shaded regions in the Fig. 1) by counting the

number of base pairs in common between two regions or calculated negative when away from the ends of the two regions. An illustration of the algorithm can be found in Figure 1. There are three possibilities:

i) One region is within, or completely overlaps, the other. In this case the bp overlap is simply a positive number reported by calculating the length of the smaller region that lies within the second region. If the two regions completely overlap each other then the length of either region can be reported as the bp overlap. For example, where region A lies within region B (Fig. 1-i-a), this can be identified computationally by checking if A.End is less than or equal to B.End and if A.Start is greater than or equal to B.Start. The region length of A can be calculated by the SQL pseudocode extract given below:

```
WHEN A.End ≤ B.END AND A.Start ≥ B.Start
THEN (A.End - A.Start)
```

Conversely, when region B lies within region A (Figure 1-i-b) or completely overlaps, this can be confirmed by checking whether B.End is less than or equal to A.End and if B.Start is greater than or equal to A.Start. The region length of B can then be calculated:

```
WHEN B.End ≤ A.END AND B.Start ≥ A.Start
THEN (B.End - B.Start)
```

ii) Region A is located on the left side of region B. In this possibility the two regions may share bases in common (Figure 1-ii-a) or can be completely away from each other (Figure 1-ii-b). Computationally

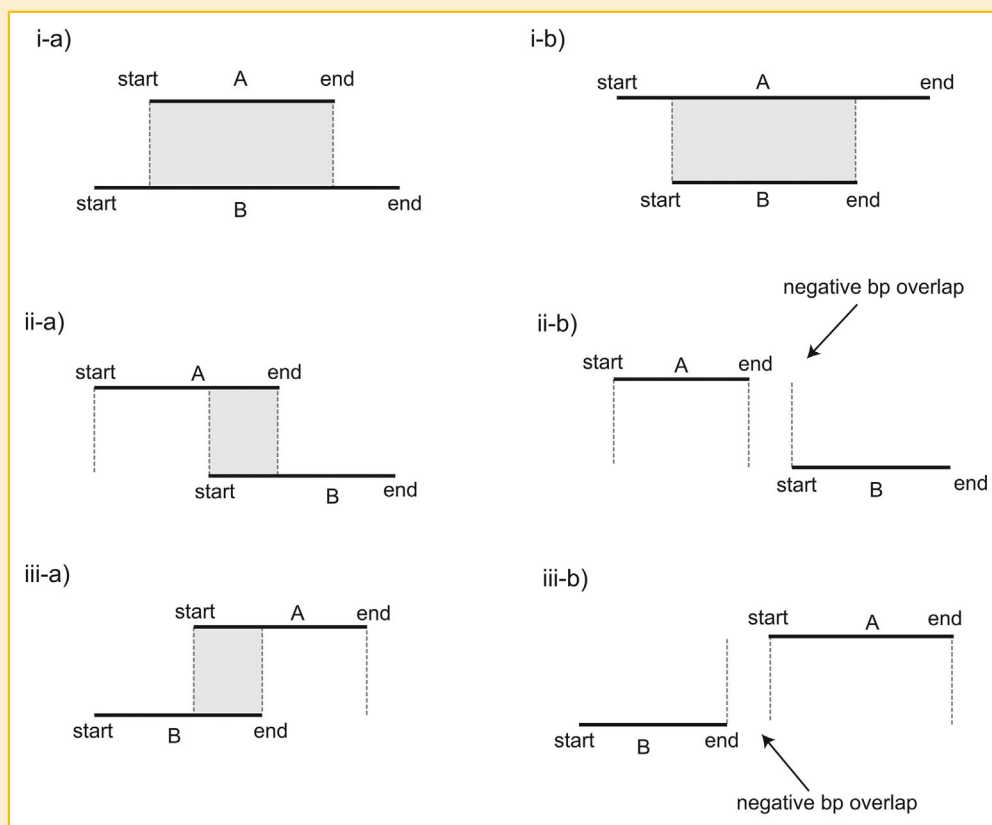


Fig.1. Various possible relative positions of the two genomic regions. (i) One genomic region is completely within the other. (ii) The overlapping or non-overlapping region A is on the left side of the region B. (iii) The overlapping or non-overlapping region A is on the right side of the region B.

this is verified by checking whether A.End is less than or equal to B.End and A.Start is less than or equal to B.Start. The bp overlap is calculated by subtracting B.Start from A.End:

```
WHEN A.End ≤ B.End AND A.Start ≤ B.Start  
THEN (A.End – B.Start)
```

Using the above calculation, for the first situation (Fig. 1-ii-a) the bp overlap will be reported as a positive integer. For the second situation (Fig. 1-ii-b) when no common bases exist between the two regions, the bp overlap is returned as a negative number. This is because the B.Start coordinate is greater than A.End therefore (A.End – B.Start) will be a negative number.

iii) Region A is located on the right side of region B. As above, the two regions could overlap or can be apart without intersecting each other. This is verified by checking whether A.End is greater than or equal to B.End and if A.Start is greater than or equal to B.Start. The bp overlap is calculated by subtracting A.Start from B.End:

```
WHEN A.End ≥ B.End AND A.Start ≥ B.Start  
THEN (B.End – A.Start)
```

Similar to the above case the overlapping regions (Fig. 1-iii-a) will have positive bp overlap and non-overlapping regions will have negative bp overlap (Fig. 1-iii-b).

I also calculated the distance between the centres of two regions, referred to as 'centre distance'. Sometimes a small base pair overlap of very long regions does not make any biological sense so overlap analysis may be required to limit to a certain distance from the centre of two regions. This is also useful in extracting regions that do not overlap but are very close to each other. To calculate the distance, the centres of the two regions are determined and then the absolute (positive) distance between the centres is calculated by the following SQL code extract.

```
abs ((A.End + A.Start)/2 – (B.End + B.Start)/2)
```

The full code for the algorithm has been provided in Supplementary File 1.

BENCHMARKING FOR INSERTION OF GENOMIC REGIONS

RegMap script randomly generates region data between the specified lower and upper range which is temporarily saved in memory and then saved in the database in a single transaction. This technique was faster for both databases, since each time an insert statement was executed against the database there were transaction overheads. Therefore, generating and saving regions one by one was much slower than saving all the data at once.

I tested the performance by generating 5 K, 10 K, 20 K, 40 K and 80 K regions and identified that PostgreSQL's generation of random regions and insertion was much faster than MySQL in both InnoDB and MyISAM storage engines. MySQL's insertion of regions was dramatically slower and the time taken was almost double by doubling the number of regions inserted (Fig. 2). MySQL-InnoDB performed slightly better than MySQL-MyISAM, therefore, in Figure 2 performance of MySQL-InnoDB is reported. PostgreSQL (RegMap) generated and saved 5 K random regions in 1 s as compared to 219 s in MySQL-InnoDB and 237 s in MySQL-MyISAM, indicating that MySQL was ~220 times slower. This difference

dramatically increases for a much larger number of regions. For generating 80 K random genomic regions PostgreSQL took 4 s as compared to 3,596 s in MySQL-InnoDB and 3,680 s in MySQL-MyISAM.

In addition, the write performance was tested by importing the 1005 files consisting of 23,827,431 real genomic regions, collected from previous studies, into both databases using bulk import statements of the databases. PostgreSQL *COPY* statement while MySQL *LOAD DATA INFILE* statement was used for this purpose. I performed the import of each file in three steps: i) data was imported into a staging table, ii) data was copied across the production table while assigning a unique id, and iii) the staging table was emptied. This procedure was adopted because in reality the imported data usually needs to be processed and assigned a unique identity in order to link to other tables. PostgreSQL performed >5 times better than MySQL, PostgreSQL took ~445 s compared to ~2,940 s in MySQL-InnoDB and 2,460 s in MySQL-MyISAM. The actual import script is also provided in the Supplementary File 1.

Data upload performance is critical for bioinformatics servers where many users insert a large amount of data at once. Therefore this benchmark identified that PostgreSQL inserts and imports data much faster than MySQL.

BENCHMARKING FOR IDENTIFICATION OF OVERLAPPING REGIONS

I further investigated the performance of reporting intersecting or overlapping regions using RegMap and using the database built-in functions in each database. The two databases have built-in functions that can be used to identify intersecting lines. Since these built-in functions are usually used in geographical (spatial) mapping software I subsequently refer to the built-in functions as Geo functions.

PostgreSQL's performance was again outstanding in finding overlapping genomic regions compared to MySQL (Fig. 3). RegMap in PostgreSQL took 134 s to report intersecting regions for two datasets of 80 K regions each, and 257 s using PostgreSQL-Geo function. MySQL performance was tested using InnoDB and MyISAM storage engines. MySQL-MyISAM performed poorly compared to InnoDB, however, both engines demonstrated inferior performance as compared to PostgreSQL. For example, when two datasets of 80 K regions were tested for overlaps using RegMAP, MySQL-InnoDB took 1,119 s, and MySQL-MyISAM took 1,150 s. Therefore, for simplicity reasons, I presented the data for MySQL-InnoDB in Figure 3.

Applying various indexes on these regions did not improve performance in PostgreSQL while it had a negative effect in MySQL for both engines (InnoDB & MyISAM). I performed these tests on different computers and obtained slightly different timings, however, the overall outcome remained the same which was that PostgreSQL performance in identifying overlapping regions was much better than MySQL.

Since RegMap identifies overlapping regions by calculating 'bp overlap' for each region, I finally concluded that queries that require extensive calculation of mathematical operations perform much better in PostgreSQL.

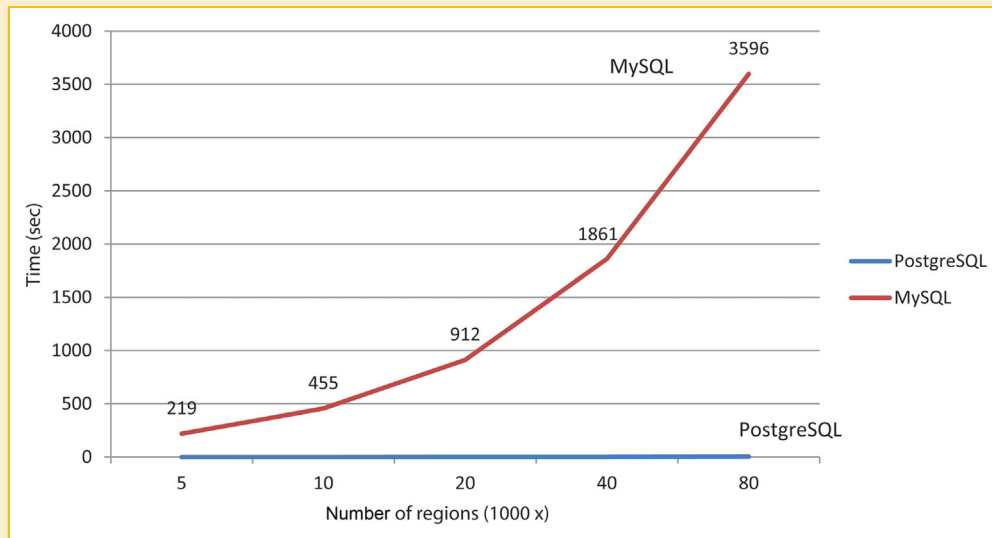


Fig.2. Comparison of region insertion performance. For simplicity, MySQL times shown are for InnoDB storage engine as MyISAM did not perform as well.

SEARCHING AND RETRIEVING REGIONS

PostgreSQL was slightly better at performing a search of genomic regions than MySQL. The knowledge base of ~24 million genomic regions was searched for erroneous regions with a start coordinate less than 0 or an end coordinate less than start. PostgreSQL identified 10 erroneous regions in ~5s while MySQL-InnoDB found the same erroneous regions in ~21 s and MySQL-MyISAM in 6 s. Implementing various types of indexes on chromosome start and end fields did not

improve performance for this query. However, searching for specific regions within a certain distance of a gene was instant in all databases. For example, searching regions within 100,000 upstream/downstream of the transcription start site of MYC gene (chr8:128748314) returned results in 3–5s in all databases which was further reduced to ~1 s by implementing an index. Therefore I concluded that the general searching capability of PostgreSQL and MySQL is similar. The Queries are provided in the Supplementary File 1.

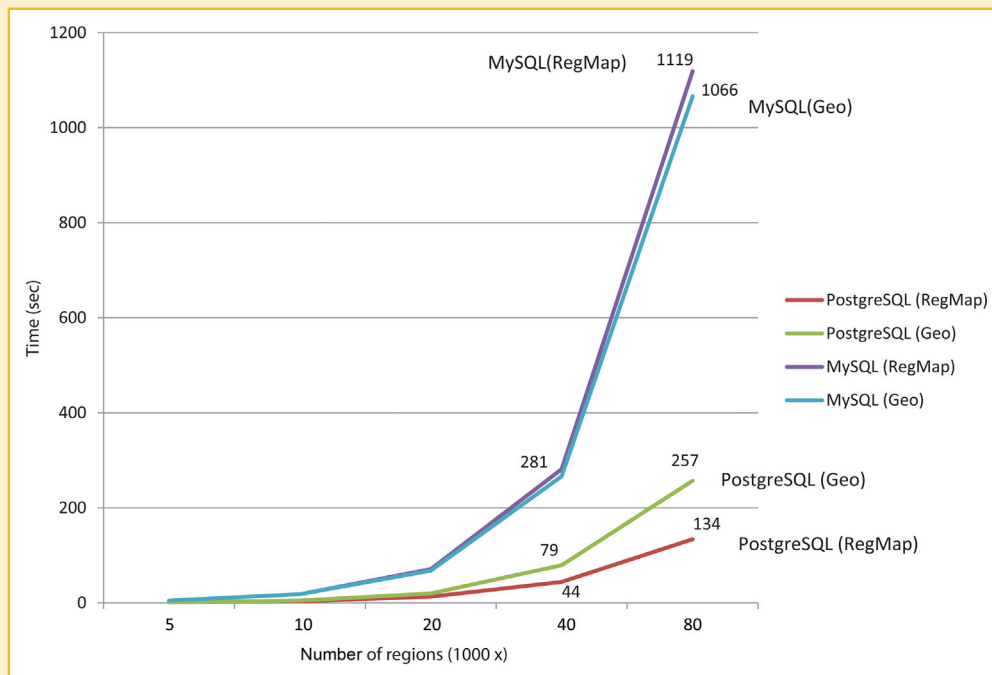


Fig.3. Comparison of performance for identifying overlapping regions using RegMap and Geo functions. For simplicity, InnoDB times are shown for MySQL, as the MyISAM storage engine in MySQL did not perform as well.

ADVANTAGES OF RegMap OVER GEO FUNCTIONS

The RegMap algorithm not only outperformed the Geo functions, in addition, it provides extended functionality that the Geo functions do not provide. For example, Geo functions only return a Boolean (true/false) value if the queried regions intersect or not. On the other hand, RegMap reports the number of bases common in the two regions or away from each other. Therefore it is easy to limit results based on the minimum number of bases that must overlap. It also provides the ability to restrict results based on the distance from the centre of regions; this is useful in returning regions that do not share common bases, but are present in close proximity within a specified distance. For example the SQL query *select * from vwregions where bpooverlap<1 and centredistance<1000*; will return regions that do not overlap however their centres are within a distance of 1000 bases. This type of analysis is important in identifying partner factors that bind on DNA in close proximity to each other without overlapping.

MINING KNOWLEDGE BASE

The Supplementary File 2 contains overlapping results of >1000 datasets of transcription factor binding sites and histone marks with information about cell line, treatment condition (if there is any), total number of regions, the number of overlapping regions and its percentage found against other datasets (Supplementary File 2). The results are spread across four spreadsheets based on the aligned reference genome i.e. hg 19, hg 18, mm 9 or mm 8. Researchers can easily filter records based on restricting values in each field and then sorting on 'Percentage Overlaps' to find out the most or least interacting dataset. There are links provided to the raw data and to the publications.

This useful knowledge base can help in developing new hypotheses that can further be tested and analysed in the wet lab. For example, using the knowledge base a novel cis-regulatory interaction between estrogen receptor alpha (ER α) and progesterin receptor (PR) was identified [Khushi et al., 2014a]. I observed that among all factors in hg 18 assembly SA1, Rad21 and CTCF binding locations were comparatively conserved in MCF7 (breast adenocarcinoma cell line) when compared to their binding locations in H1 hESC (human embryonic stem cell line). These factors targeted similar genomic locations in the two distinct cell-lines, despite previous reports describing the binding pattern of various factors to be cell specific [The ENCODE Project Consortium, 2012; Wang et al., 2012].

In the HepG2 (liver hepatocellular) cell line, I identified that HNF4G preferentially binds (6633/6839, 97%) to H3K4Me1 (enhancer) regions, and the majority of HNF4G binding sites (4244/6839, ~62%) were also found overlapping with STAG1 binding sites (83080 regions). The statistical significance of overlapping of HNF4G is further analysed in BiSA [Khushi et al., 2014b] which revealed a statistically significant overlap correlation value of 0.65. As previously described, the overlap correlation value greater than 0.5 shows a statistical significant overlap of a query factor [Khushi et al., 2014b]. HNF4G is an orphan nuclear receptor whose ligand and function has not been fully understood, however recent studies have shown HNF4G overexpression to induce growth of cancer tissue [Okegawa et al., 2013; Yang et al., 2014]. On the

other hand, STAG1 (Stromal Antigen 1), also known as SA1, is one of the four subunits of the cohesin complex [Losada, 2014]. Cohesin has important roles in transcription regulation, DNA repair, chromosome condensation, homolog pairing, etc. [Mehta et al., 2013; Losada, 2014]. Therefore, statistical significant overlap of HNF4G with STAG1 indicates an important underlying biology which could be further explored in laboratory.

DISCUSSION

Various databases are heavily used in biomedical and cellular biochemistry, therefore researchers would benefit from knowing which database product performs better for a specific type of data. Benchmarking software products also helps vendors to improve their products. Various other benchmarks for database systems exist and it is acknowledged that development and adoption of benchmarks advances research in a research area [Sim et al., 2003; Arslan and Yilmazel, 2008; Bose et al., 2009; Aniba et al., 2010; Ray et al., 2011; Venema et al., 2013]. For example, Ray et al. [2011] benchmarked databases for spatial data, whereas, Xu et al. [2012] benchmarked databases for moving objects data. However no benchmarking effort exists on database performance for genomic region operations. Therefore RegMap, being natively written in SQL and adaptable for any SQL-based database, will advance research in this field and will provide a baseline mark for future algorithms.

Many bioinformatics analyses produce a large number of variant files. Usually detailed information about factor, cell-line, condition, peak-calling or analysis parameters used are recorded as part of file names or kept separate which makes it difficult to manage such information for a large scale study. Databases provide a more effective way of managing curation, annotation, sorting and relationships among data. RegMap, being a SQL-based algorithm, can be integrated into any language as most languages provide API (application programming interface) to connect to SQL-based databases. SQL's simple syntax is also easy for biologists to learn. There are a number of tools that are in use by the research community to operate on genomic regions, for example BEDTools [Quinlan and Hall, 2010], Pybedtools [Dale et al., 2011], GenomicTools [Tsirigos et al., 2012], and BEDOPS Tools [Neph et al., 2012]. All of these tools are designed to operate on files and integration of these tools in other languages is usually difficult. Tabix [Li, 2011] is another efficient tool that is usually used to extract specific regions from large files. The database capability of searching specific genomic regions was equivalent to Tabix. Both databases, when searching a table with ~24 million real genomic regions, returned results in ~1 s for regions that were within 100 K of transcription start site of MYC gene. There are only a few tools which provide an easy interface to other languages, such as Pybedtools which provides Python interface, and GROK and GenomicTools which provide C++ API (application programming interface) to C/C++ programmers. There are a few GUI (Graphical User Interface) tools such as Cisgenome [Ji et al., 2011], Galaxy [Goecks et al., 2010], and the UCSC table browser [Karolchik et al., 2004] which provide very basic genomic operation analysis options. However, there is no algorithm

available that performs genomic region operations natively in a relational database system. Therefore direct comparison of the performance between RegMap algorithm and other tools that work on files is not appropriate.

RegMap, being an SQL based algorithm, is easy to apply on an unlimited number of datasets. The algorithm was applied to ~1000 datasets of transcription factor binding sites and epigenetic marks and an easy navigate-able Excel file was generated. This data can be used to develop new hypotheses such as identification of novel biochemical partners of a factor or factor's binding influenced by a histone mark. Once an interesting interaction is found, actual genomic locations could be studied using tools such as BiSA [Khushi et al., 2014b]. Using the knowledge base a novel interaction between HNF4G and cohesion subunit STAG1/SA1 was identified. The majority of HNF4G binding sites overlapped STAG1 and this overlap was statistically significant suggesting an important biochemical partnership on a specific subset of genomic regions.

The performance of proprietary databases such as Oracle or Microsoft SQL Server was not reported because of their licensing restrictions, however, using the algorithm researchers/institutions can benchmark the suitability of either product for their own use.

I acknowledge that in other computational infrastructure database performance could vary, however, I have done rigorous testing on different machines to conclude that similar relative results would be obtained. The results also revealed that there is a great deal of room present to improve the database built-in functions that are used to find intersecting geometrical shapes. In other fields such as geo mapping application it is usually not required to find the thousands of intersecting features. However genomic studies deal with a large amount of data. With increased use of databases in genomic applications, there is a need for database functions to be improved for genomic operations. Therefore it is proposed here that 'genomic region' data-type in all databases should be implemented.

In summary, RegMap is an open source database-driven algorithm used to find overlapping/non-overlapping regions, and results can be limited by the number of bases in common or maximum distance between the centres of two sets. Using the algorithm I benchmarked performance of two widely used open source databases, PostgreSQL and MySQL. The benchmark revealed that PostgreSQL performs much better in identifying overlapping genomic regions. Data upload/import function of PostgreSQL was also better than MySQL. Data upload performance is critical for bioinformatics facility servers where many users insert a large amount of data simultaneously. Using the algorithm the overlapping of >1000 datasets of transcription factor binding sites and histone modifications were calculated and identified that HNF4G binding significantly overlaps with cohesion subunit STAG1/SA1 binding on DNA.

ACKNOWLEDGEMENTS

The author was supported by an Australian Postgraduate Award (APA) and Westmead Medical Research Foundation (WMRF) Top-Up scholarship. The author would like to deeply thank Professor Christine L. Clarke and Dr. J. Dinny Graham for their supervision during PhD candidature and to Dr. S. Cunningham for correcting some grammatical errors.

REFERENCES

- Aniba MR, Poch O, Thompson JD. 2010. Issues in bioinformatics benchmarking: The case study of multiple sequence alignment. *Nucleic Acids Res* 38:7353–7363.
- Arslan A, Yilmazel O. 2008. A comparison of relational databases and information retrieval libraries on turkish text retrieval. *International Conference on Natural Language Processing and Knowledge Engineering, 2008 (NLP-KE '08)*. Washington, DC: IEEE. pp 1–8.
- Bose S, Mishra P, Sethuraman P, Taheri R. 2009. Benchmarking database performance in a virtual environment. In: Nambiar R, Poess M, editors. *Performance Evaluation and Benchmarking*. Berlin, Heidelberg: Springer. pp 167–182.
- Dale RK, Pedersen BS, Quinlan AR. 2011. 2011. Pybedtools: A flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* 27(24):3423–3424.
- Goecks J, Nekrutenko A, Taylor J. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 11:R86.
- Ji H, Jiang H, Ma W, Wong WH. 2011. Using CisGenome to analyze ChIP-chip and ChIP-seq data. *Curr Protoc Bioinformatics Chapter 2:Unit2*. 13.
- Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32: D493–D496.
- Khushi M, Clarke CL, Graham JD. 2014. Bioinformatic analysis of cis-regulatory interactions between progesterone and estrogen receptors in breast cancer. *PeerJ* 2:e654.
- Khushi M, Liddle C, Clarke CL, Graham JD. 2014. Binding sites analyser (BiSA): Software for genomic binding sites archiving and overlap analysis. *PLoS One* 9:e87301.
- Li H. 2011. Tabix: Fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* 27:718–719.
- Losada A. 2014. Cohesin in cancer: Chromosome segregation and beyond. *Nat Rev Cancer* 14:389–393.
- Mehta GD, Kumar R, Srivastava S, Ghosh SK. 2013. Cohesin: Functions beyond sister chromatid cohesion. *FEBS Lett* 587:2299–2312.
- Meyer CA, Tang Q, Liu XS. 2012. Minireview: Applications of next-generation sequencing on studies of nuclear receptor regulation and function. *Mol Endocrinol* 26:1651–1659.
- Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, Sandstrom R, Humbert R, Stamatoyannopoulos JA. 2012. BEDOPS: High-performance genomic feature operations. *Bioinformatics* 28:1919–1920.
- Okegawa T, Ushio K, Imai M, Morimoto M, Hara T. 2013. Orphan nuclear receptor HNF4G promotes bladder cancer growth and invasion through the regulation of the hyaluronan synthase gene. *Oncogenesis* 2:e58.
- Padilla A, Hawkins T. 2011. Database optimization. In: *Pro PHP Application Performance*. New York: press. pp 189–208.
- Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.
- Ray S, Simion B, Brown AD. 2011. Jackpine: A benchmark to evaluate spatial database performance. *IEEE 27th International Conference on Data Engineering (ICDE)*. Washington, DC: IEEE Computer Society. pp 1139–1150.
- Sheldon R, Moes G. 2005. *Beginning MySQL*. Indianapolis, IN: Wiley.
- Sim SE, Easterbrook S, Holt RC. 2003. Using benchmarking to advance research: A challenge to software engineering. *Proceedings of the 25th International Conference on Software Engineering*. IEEE Computer Society. pp 74–83.
- The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.

Tsirigos A, Haiminen N, Bilal E, Utro F. 2012. GenomicTools: A computational platform for developing high-throughput analytics in genomics. *Bioinformatics* 28:282–283.

Venema V, Mestre O, Aguilar E, Auer I, Guijarro J, Domonkos P, Vertacnik G, Szentimrey T, Stepanek P, Zahradnicek P. 2013. Benchmarking homogenization algorithms for monthly data. *Proceedings of the Ninth International Temperature Symposium. AIP Conference Proceedings*. p 1060.

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

Xu J, Güuting RH. 2012. GMOBench: A Benchmark for Generic Moving Objects. *Informatik-Report* 362, Fernuniversität Hagen.

Yang CS, Chang KY, Rana TM. 2014. Genome-wide functional analysis reveals factors needed at the transition steps of induced reprogramming. *Cell Rep* 8:327–337.

Zammataro L, DeMolfetta R, Bucci G, Ceol A, Muller H. 2014. Annotate-GenomicRegions: A web application. *BMC Bioinformatics* 15:S8.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site.

Binding Sites Analyser (BiSA): Software for Genomic Binding Sites Archiving and Overlap Analysis

Matloob Khushi^{1,3*}, Christopher Liddle², Christine L. Clarke¹, J. Dinny Graham¹

1 Westmead Institute for Cancer Research, Sydney Medical School, University of Sydney and the Westmead Millennium Institute, Westmead, New South Wales, Australia, **2** Storr Liver Unit, Sydney Medical School, University of Sydney and the Westmead Millennium Institute, Westmead, New South Wales, Australia, **3** Australian Breast Cancer Tissue Bank, Sydney Medical School, University of Sydney and the Westmead Millennium Institute, Westmead, New South Wales, Australia

Abstract

Genome-wide mapping of transcription factor binding and histone modification reveals complex patterns of interactions. Identifying overlaps in binding patterns by different factors is a major objective of genomic studies, but existing methods to archive large numbers of datasets in a personalised database lack sophistication and utility. Therefore we have developed transcription factor DNA binding site analyser software (BiSA), for archiving of binding regions and easy identification of overlap with or proximity to other regions of interest. Analysis results can be restricted by chromosome or base pair overlap between regions or maximum distance between binding peaks. BiSA is capable of reporting overlapping regions that share common base pairs; regions that are nearby; regions that are not overlapping; and average region sizes. BiSA can identify genes located near binding regions of interest, genomic features near a gene or locus of interest and statistical significance of overlapping regions can also be reported. Overlapping results can be visualized as Venn diagrams. A major strength of BiSA is that it is supported by a comprehensive database of publicly available transcription factor binding sites and histone modifications, which can be directly compared to user data. The documentation and source code are available on <http://bisa.sourceforge.net>

Citation: Khushi M, Liddle C, Clarke CL, Graham JD (2014) Binding Sites Analyser (BiSA): Software for Genomic Binding Sites Archiving and Overlap Analysis. PLoS ONE 9(2): e87301. doi:10.1371/journal.pone.0087301

Editor: Manuela Helmer-Citterich, University of Rome Tor Vergata, Italy

Received: September 8, 2013; **Accepted:** December 24, 2013; **Published:** February 12, 2014

Copyright: © 2014 Khushi et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: MK is supported by Australian Postgraduate Award (APA) and Westmead Medical Research Foundation (WMRF) Top-Up Grant. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mkhushi@uni.sydney.edu.au

Introduction

The recent revolution in whole genome census approaches has seen an exponential increase in available data sets describing genomic features, such as transcription factor (TF) binding sites and histone modifications. Recent studies have revealed that there are often overlaps and co-association between transcription factors at binding sites [1,2] and identifying relationships, such as overlaps in genomic features, has become a fundamental biological research tool [3]. Moreover, the existence of a wealth of published data sets now presents unprecedented opportunities for data mining in large databases of archived genomic data.

Existing methods of finding overlaps such as BEDTools, UCSC Table Browser, Homer or Segor [4,5,6] are limited in functionality for simultaneous comparison to multiple archived data sets. Moreover, few tools provide a simple interface that can be easily implemented by biologists with limited computing skills.

To address these challenges, we have developed BiSA, which is pre-loaded with transcription factor binding sites and histone modification locations, for a range of cell-types and conditions, reported in previous ChIP-chip and ChIP-Seq studies. BiSA allows the investigator to analyse overlapping or non-overlapping regions, to visualise results by Venn diagram, and to identify the genes located near to regions under study. BiSA is controlled through a user-friendly graphical user interface (GUI), installed on a Windows environment or embedded in the Galaxy web-based high throughput genomic analysis tool. Both options maximise the

ease of use of this powerful tool for molecular biologists, who may lack the necessary computing skills required to use alternate approaches.

Methods

BiSA employs a rational database management system-based architecture to archive unlimited numbers of binding datasets in a very flexible and convenient format. BiSA is developed in C# and SQL Server 2008 for the Windows environment, while Python and PostgreSQL have been used to develop a Linux version that runs under the Galaxy [7] web-based environment. We have used Google Charts to generate Venn diagrams. To calculate common sections on Venn diagrams for three datasets, BiSA first extracts overlapping regions of two datasets and then overlaps the result with the third dataset.

There are three main steps of installation of BiSA for Windows: i) installation of the MS SQL Server database engine. BiSA is fully compatible with the free Express Edition of SQL Server [8], ii) downloading and restoring the BiSA database file using SQL Server Management Studio, and iii) linking the front-end application to the database. Detailed step-by-step installation instructions with screenshots for Windows and Linux environments are available at the project website <http://bisa.sourceforge.net/>. We will periodically update the database to include datasets from the latest published studies.

Figure 1. Import Datasets to Knowledge Base (KB). This step is optional and users can study data already saved in the KB, without importing datasets. In this step, the user can upload their own transcription factor DNA binding sites or histone modification locations, usually as BED or GFF peak files. If the file extension is other than BED or GFF, BiSA prompts the user to choose the right format. It is important to specify a Reference Genome (encircled), for instance hg18/hg19 for human or mm9/mm8 for mouse. BiSA will only allow comparisons between datasets of the same reference genome.

doi:10.1371/journal.pone.0087301.g001

The BiSA database schema is straightforward. All information about a dataset such as factor label, cell line and condition are saved in the 'kbdetails' table while the genomic region data are saved in the 'kbsites' table and linked to the 'kbdetails' table by an identity. There are four gene annotation tables covering the human hg19 and hg18, and mouse mm9 and mm8 genome assemblies. In addition to the options provided by the graphical user interface (GUI), a user can analyse data via SQL Server Management Studio by structured query language (SQL). SQL is an accessible database interrogation language, and simple SQL statements can be used to analyse data, for example:

1. Erroneous reporting of an end coordinate smaller than the start coordinate for a genomic region can be discovered by the SQL statement: `SELECT * FROM kbsites WHERE [end]<start`. Running this query on the BiSA KB, interrogated ~18 million regions in less than 1 second and discovered one dataset where this error was present.
2. The gene annotation tables in BiSA contain gene names and symbols, NCBI accession IDs, chromosome, strand, and the coordinates of transcription start site (TSS), end site (TES),

coding sequence (CDS) and exon positions. To return all annotation details in the hg19 assembly for the breast cancer susceptibility gene BRCA1, an investigator could use the query: `SELECT * FROM Annotation_hg19 WHERE gene_id = 'BRCA1'`

3. To report all genomic features within 100 kb upstream or downstream of the BRCA1 TSS. The user could then use the coordinates returned in example 2, in the SQL query: `SELECT * FROM kbsites WHERE chr = 'chr17' AND start >= 41096311 AND start <= 41296311`
4. Average region length in dataset ID 10 can be retrieved by the SQL statement: `SELECT AVG([end] - start) FROM kbsites WHERE kbid = 10`

Initially we have populated the BiSA database with ~600 datasets of transcription factor binding sites and histone modifications amounting to approximately 18 million genomic regions. The data have been collected from previously published studies deposited on the publishing journals' websites, Gene Expression Omnibus (GEO), European Bioinformatics Institute (EBI),

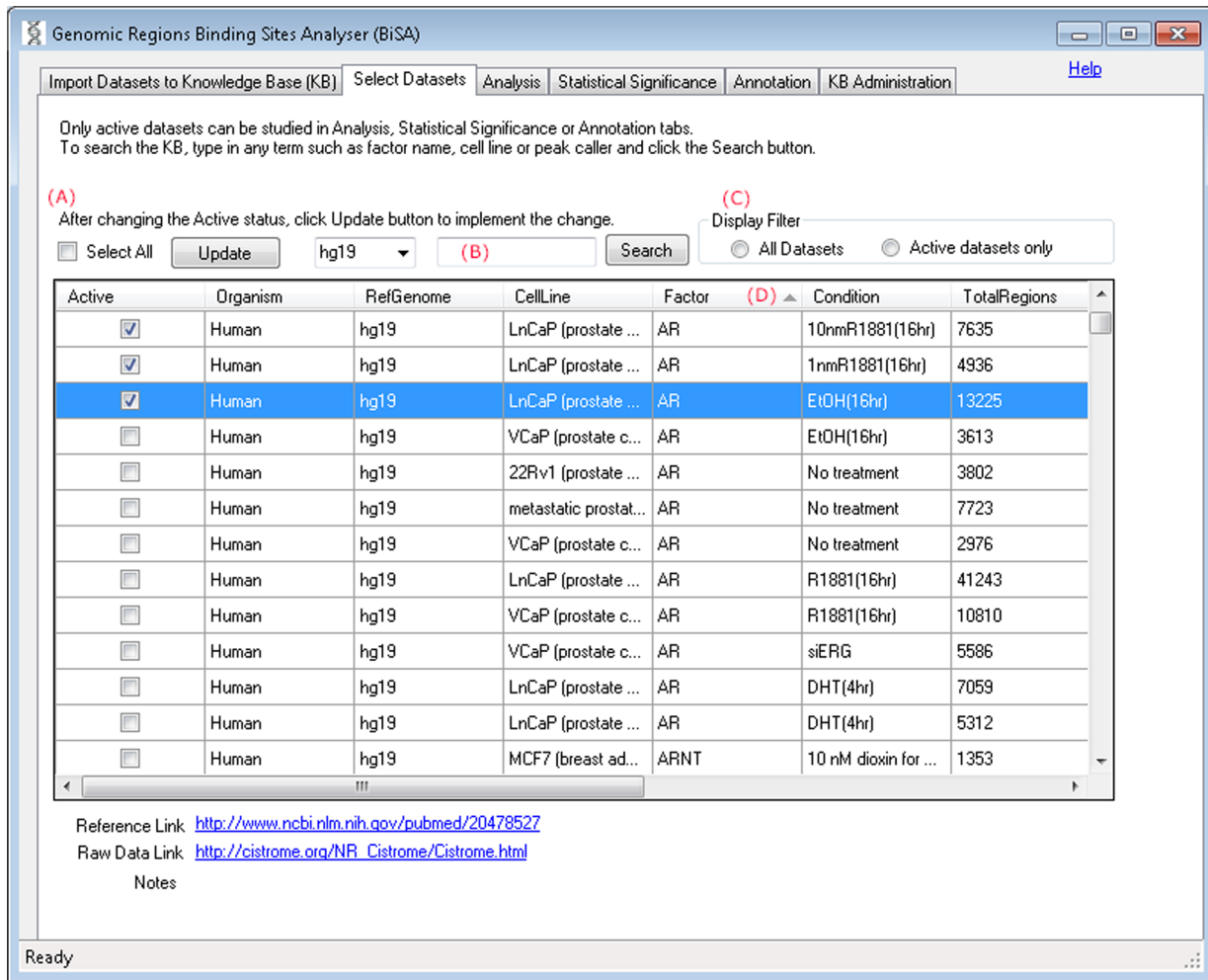


Figure 2. Select Datasets. This tab displays a list of all datasets in the KB, populated by default or as a consequence of uploading in Step-1. Clicking on the text of any row displays the reference link of the article, raw data link and notes, if any, below the table. Website addresses are hyperlinked to the websites/articles from where the data are obtained. (A) Changing the Active ticks and clicking on the Update button implements the selection. (B) Users can search the KB by organism, cell line, factor label, reference genome or peak caller. doi:10.1371/journal.pone.0087301.g002

Cistrome Project [9] and some datasets are collected directly from the authors. The source of the data and additional comments, if there are any, are recorded in the kbdetails table. Addition of more datasets is a straightforward process. We refer to the sum of these datasets as the Knowledge Base (KB). Users can expand on an existing KB or build their own KB. The KB, currently, comprises human (hg18 & hg19 build) and mouse (mm8 and mm9) assemblies.

BiSA steps through the process of studying binding region interaction, annotation, statistical significance and management of datasets in seven GUI screens. An investigator can study overlapping regions by setting the minimum base pair (bp) overlap. It is also possible in BiSA to limit reported overlaps based on a maximum distance between binding peaks. This is important since binding region boundaries are highly dependent on the peak caller software and parameters used.

We have implemented IntervalStats [10] in BiSA to test the statistical significance of overlap between two dataset. IntervalStats is a command line tool written mainly for the Unix environment. Therefore, we used the MinGW toolkit [11] to compile it for the Windows environment. The BiSA for Windows download package

includes an IntervalStats executable file and dependent DLL libraries, however, the tool runs independently of BiSA. When the IntervalStats tool is executed through the BiSA GUI, the datasets under study are saved in the 'data' subfolder and the files are passed to the IntervalStats tool. During the execution of the statistical tool the terminal window stays open to show the messages from the tool. IntervalStats calculates a p-value for each region in a query dataset against the nearest region from a reference dataset. A defined domain dataset, representing the linespace of possible interval locations, acts as a background to the statistical test and can be restricted to specific locations, such as promoter proximal regions, to take into account known biases in binding site detection. In the simplest case, the domain comprises the entire genome. We have populated BiSA with a number of domain files such as promoter regions within 10 kb of a TSS, intergenic regions and whole genome for hg19, hg18, mm9 and mm8 assemblies. Users can select one of the prepopulated domains or can specify a BED file as the domain. In addition to individual p-values for region overlap, IntervalStats returns a summary statistic, referred to as the Overlap Correlation Value, to identify the overall significance of overlap of two datasets. This

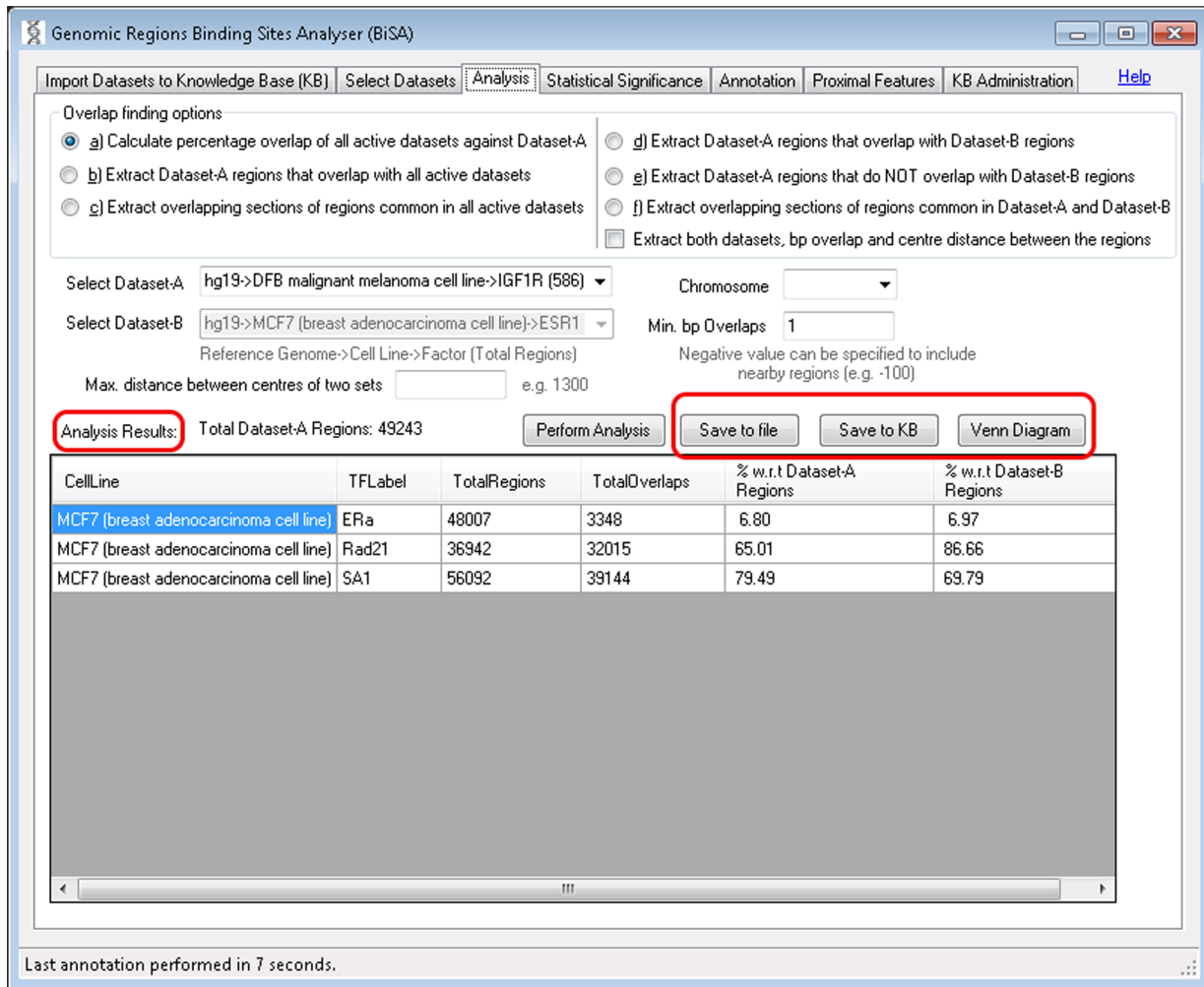


Figure 3. Analysis is the main overlap analysis tab of BiSA. BiSA offers six types of analysis: Overlap finding option a) reports overlap percentage with respect to the total Dataset-A regions and percentage with respect to the other active dataset regions. Overlapping or non-overlapping regions of Dataset-A can be extracted by options b), d) or e). Whereas, option c) or f) can be used to extract overlapping sections of regions common in all or two datasets. The results of overlap analysis type b), c), d), e) and f) can be saved back into the Knowledge Base by the 'Save to KB' button, allowing them to go into downstream analysis and independent annotation. Ticking the "Extract both datasets, bp overlap and centre distance between the regions" for options d), e) and f) displays both Dataset-A and Dataset-B regions, bp overlap and distance between the two sets. doi:10.1371/journal.pone.0087301.g003

summary statistic represents the fraction of regions in the query dataset with a p-value of overlap to the reference below a significance threshold value, and thus reflects the likely significance of overlap of the query and reference datasets. The correlation coefficient can range from 0 to 1, the closer the value to 1 the stronger the significance of overlap of two datasets. We have set the threshold p-value to 0.05, however this value can be changed in the configuration file, BiSA.exe.config if desired.

Gene annotations are obtained from the UCSC genome browser and will be updated periodically. Initially we have populated annotations for reference genomes hg18, hg19, mm8 and mm9. Custom gene definitions or transcription factor binding sites or epigenetic modifications for additional genomes for other organisms can be uploaded in the software.

Results

The BiSA Windows GUI is split across seven tabs

i) Import Datasets to Knowledge Base (KB). This is an optional step as the user can choose to analyse only data already

contained in the KB. The user browses for their dataset, which can be uploaded in tab delimited or comma delimited BED or GFF format, assigns a logical name and description for the data, and uploads to the KB (Figure 1). The first 20 lines of the data can be displayed for verification. Chromosome position is 0 indexed as in BED format. Comments or header information in the file are reported as failed records in the 'Report' section (Figure 1). If no valid data are imported in the first 50 lines, the upload fails and BiSA stops the import process. The user enters information about organism and cell line, TF and conditions, which are saved with the database record. The genome build for the genomic region coordinates must be entered during this process (Figure 1, circled) and the record will be limited in future analyses to comparison with other datasets generated in the same genome and build. Associated data and publication links can also be added at this stage.

ii) Select Datasets. This tab displays a list of all datasets in the KB including those uploaded in Step-1. Data are selected for analysis by checking the "active" box beside the relevant dataset (Figure 2A). Only data from matching reference genomes can be

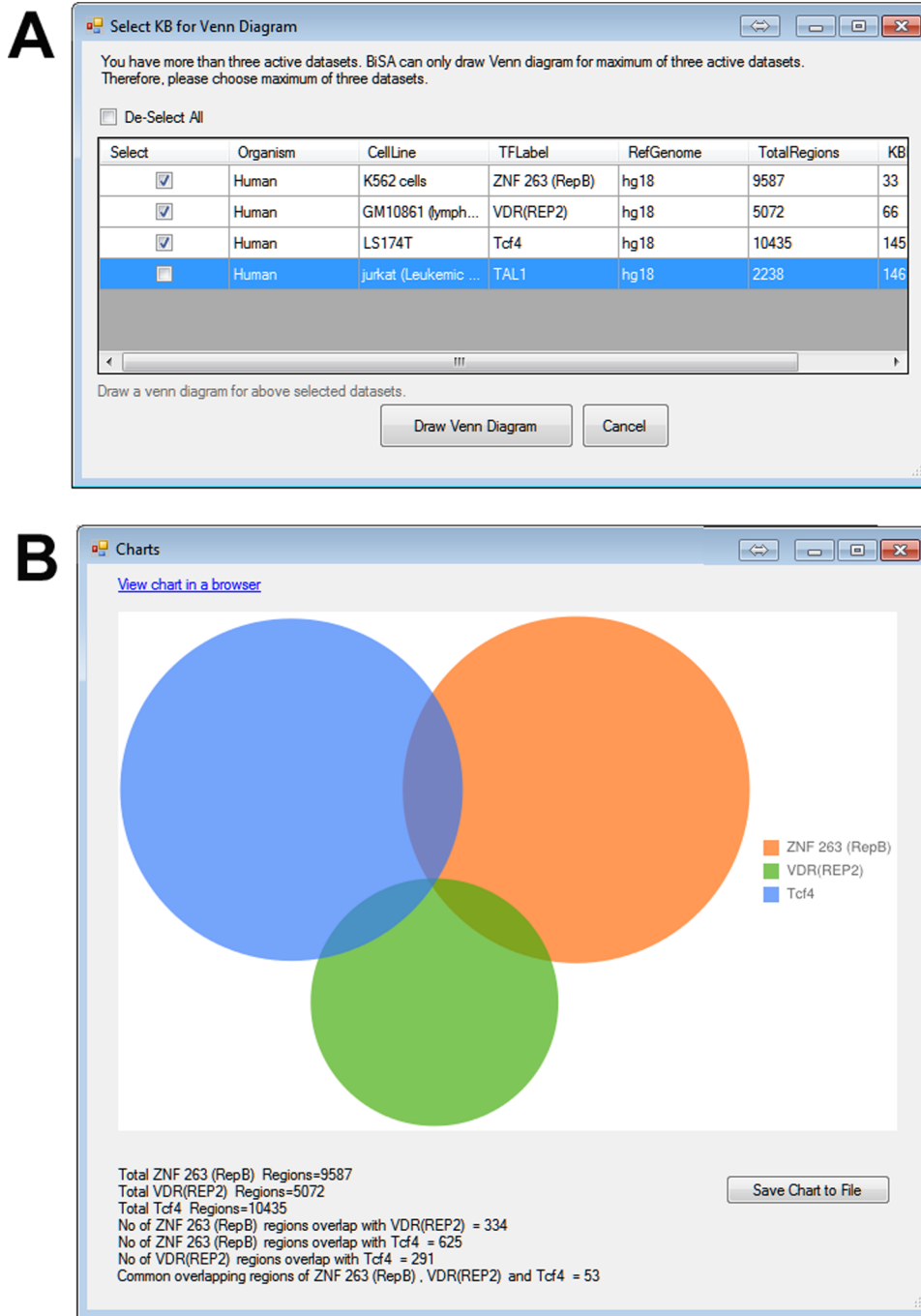


Figure 4. Venn diagrams. BiSA can cross-compare a maximum of three active KB as a Venn diagram. (A) If there are more than three active datasets then a pop-up window appears that allows the investigator to select three datasets to be analysed. (B) Google Charts is used to draw Venn diagrams. The diagram can be saved as a high quality PNG file.
doi:10.1371/journal.pone.0087301.g004

selected for analysis. A checked tick in the 'Active' column represents an active dataset that can be used in analysis in Step-3, and only active datasets can be annotated. Users can only activate datasets of matching reference genomes. To change the active status of datasets from one reference genome (e.g. hg18) to another (e.g. mm9), the user must deactivate all datasets first, which can be done by toggling on and off the 'Select All' check box and pressing the Update button. Clicking on the text of any row displays further information about the data. Website addresses are hyperlinked to

the source websites/articles for the data. After selecting datasets for analysis, clicking on the Update button activates datasets in the BiSA database. The search field (Figure 2B) allows the user to search the KB by organism, cell line, factor label, reference genome or peak caller. Only datasets that are active can be displayed by checking the 'Active datasets only' option in the 'Display Filter' (Figure 2C). Displayed data can be sorted according to any of the database fields by selecting the column heading for the field of choice (Figure 2D).

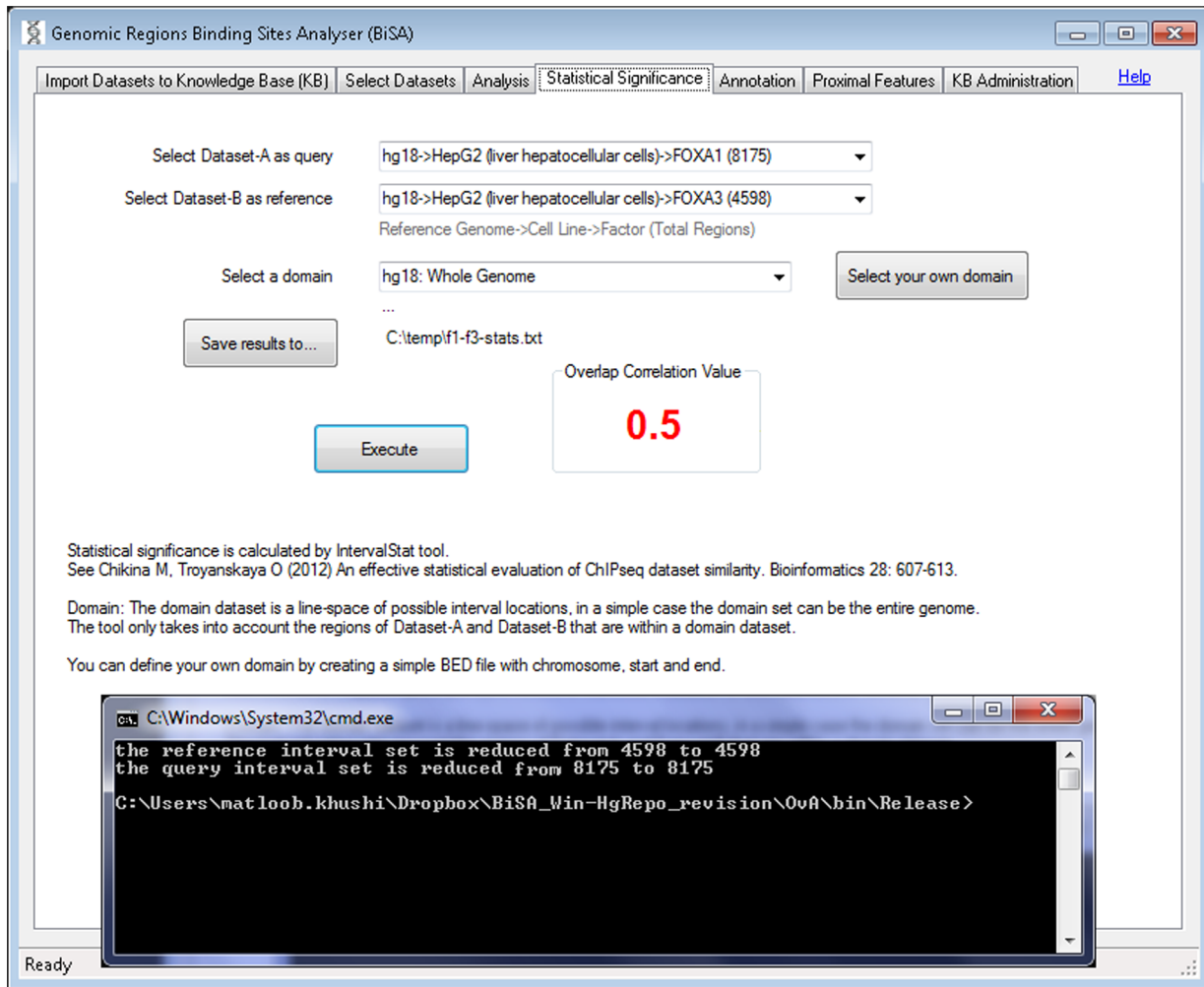


Figure 5. Statistical significance of overlapping regions. The statistical significance tab allows the user to determine the statistical significance of the extent of overlap of two sets of regions. Active datasets are loaded into two dropdown lists and the user selects one dataset as a query and the other one as a reference. Only regions of both datasets that are within the selected domain dataset are included in the calculation. Clicking the Execute button calls up a command-line window and executes the IntervalStat tool. The command-line window stays open to display the messages from the tool. When the terminal window is closed BISA calculates Overlap Correlation Value of the two datasets. doi:10.1371/journal.pone.0087301.g005

iii) Analysis. This is the main analysis screen where users can analyse active datasets. Six types of analysis are provided: a) calculate percentage overlap of all active datasets, b) extract regions that overlap with all active datasets, c) extract overlapping sections of regions common in all active datasets, d) extract regions that overlap between two selected datasets, e) extract regions that do not overlap with another selected dataset, f) extract overlapping sections of regions common in two datasets. Analysis can be restricted by chromosome. The options a), b) and c) operate on all active datasets while options d), e) and f) are designed to work on two selected datasets. Ticking the “Extract both datasets, bp overlap and centre distance between the regions” for options d), e) and f) displays both Dataset-A and Dataset-B regions, bp overlap and distance between two sets. The number of base pairs (bp) either in common in two sets (set by a positive number) or separating two sets (set by a negative number) can be specified, as can be the maximum allowed distance between the centres of two compared regions. Overlapping results can be visualized as Venn diagrams or saved to the KB or a tab delimited text file (Figure 3, circled).

All analyses require setting minimum ‘bp overlaps’, however, specifying maximum distance allowable between two binding peaks or limiting results to a chromosome is optional. A positive value for minimum bp overlap would restrict results for regions that share the specified number of common base pairs. For example, while studying TFs that compete for a specific DNA sequence or finding TFs that form a complex and bind to DNA, the minimum bp overlap can be set to 1 and maximum distance from the centre of two sets should be small, such as 50 bp. To study TFs that potentially bind close to each other but without overlap, a negative value of minimum bp overlap can be assigned to report nearby regions. For example assigning a bp overlap of -100 will report nearby regions separated by up to 100 bases, in this case, a maximum centre distance should be specified. The analysis results section is a data grid that populates the results of the performed analysis (Figure 3, circled). Results can be saved in a tab delimited text format, to allow further analysis in other software. Results can also be sorted by selecting any column heading. The Venn diagram button visualizes overlaps of a maximum of three activated datasets (Figure 3, circled). If there are more than three

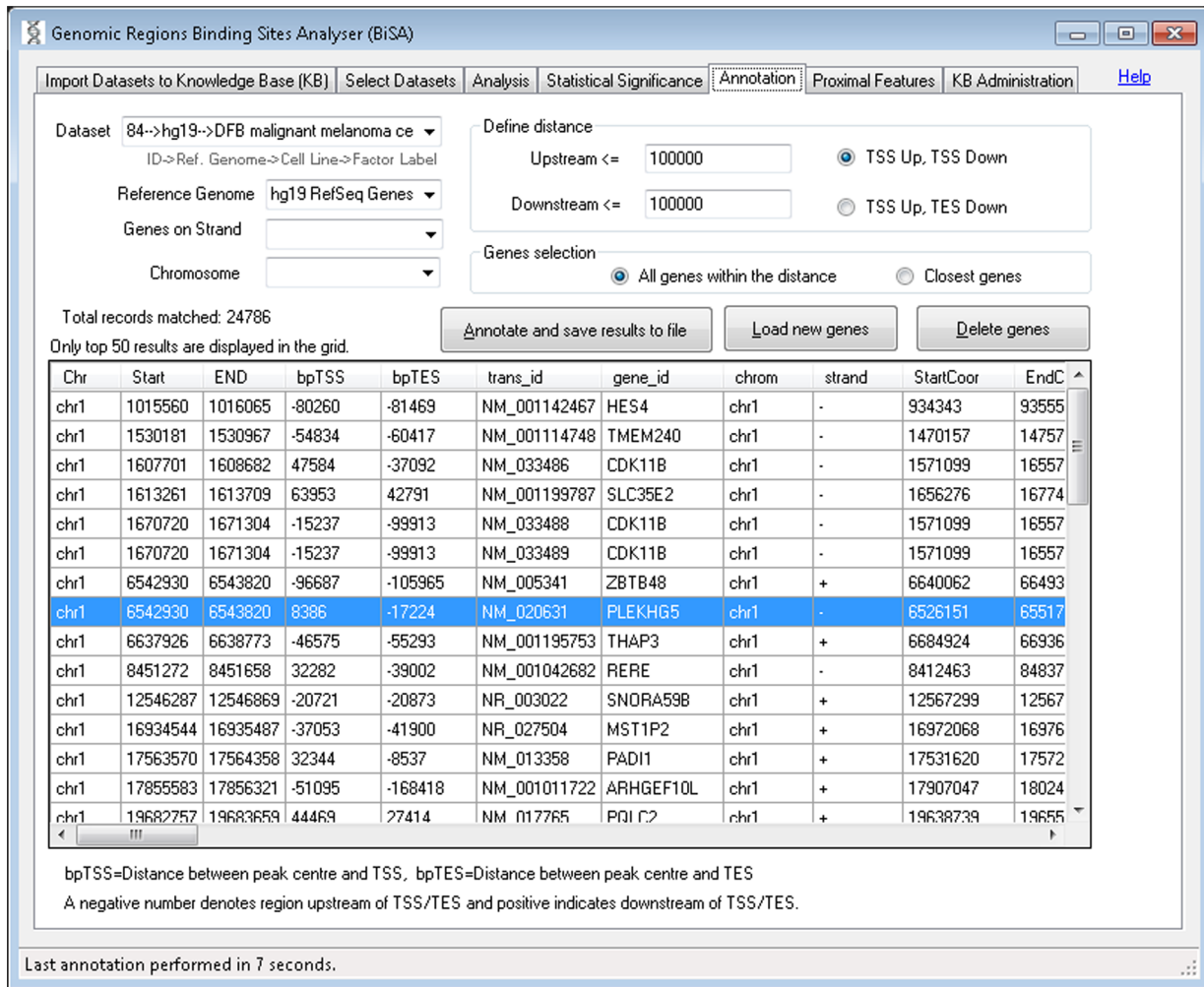


Figure 6. Gene annotation. The annotation tab allows the user to add gene information, from human and mouse reference genome assemblies, taken from the UCSC Genome Browser, to their data. This data can be saved in tab delimited text format for further analysis in other software. Annotation can be limited to a chromosome and strand. Start and End co-ordinate columns for transcript (tx) and cDNA (cds) represent the numerically lower and higher value chromosomal coordinates for genes on both strands. A negative value in the bpTSS or bpTES column indicates that the region is upstream of the annotated TSS or TES respectively. Therefore a region within a gene on the positive strand will have a negative bpTES value and a positive bpTSS value as for the region highlighted. Only the top 50 results are displayed in the grid, however, the full annotated dataset is saved in a tab delimited text file which can be opened in Excel or other spreadsheet management software for further analysis. The delete genes button allows the user to delete custom uploaded definitions. doi:10.1371/journal.pone.0087301.g006

active datasets, BiSA displays a warning (Figure 4A). Overlap statistics are displayed below the Venn diagram, which can also be saved to a file for later reference or figure preparation (Figure 4B). Overlapping or non-overlapping regions can be saved back to the KB (Figure 3) allowing them to go into downstream analysis and independent annotation.

iv) Statistical Significance. The number and location of TF binding regions discovered in a ChIP-seq experiment is influenced by experimental design, model used, sequencing depth and analysis approach. Therefore, this information is made available in as much detail as possible in BiSA, so that users can make judgements about the appropriateness of specific dataset comparisons. To determine the level of statistical significance of the degree of overlap in two datasets, the IntervalStats command line algorithm [10] is implemented in a user friendly graphical interface. Active datasets to be compared are selected via two dropdown lists (Figure 5). Users can select one dataset as a query and the other one as a reference. IntervalStats only takes into

account the regions that are within a defined domain dataset, representing the total available genomic area for binding. The results are saved as a tab delimited text file with the regions from Dataset-A (query) and Dataset-B (reference), Dataset-A region size, the distance between them and the corresponding numerator and denominator used to calculate the p-value, which is saved as the last column. Once the IntervalStats tool finishes the process and the user closes the terminal window, BiSA calculates and displays an Overlap Correlation Value as described in the Methods section, which reflects the overall significance of overlap of the two datasets.

v) Annotation. The annotation tab (Figure 6) allows the user to add nearby gene information to a selected set of binding regions. Users define maximum distances between binding peak and transcription start and end sites of nearby genes. The nearest gene per region or all genes within the designated number of bp limits will be reported. Selecting “Load new genes” (Figure 6) allows custom gene definitions for additional organisms to be

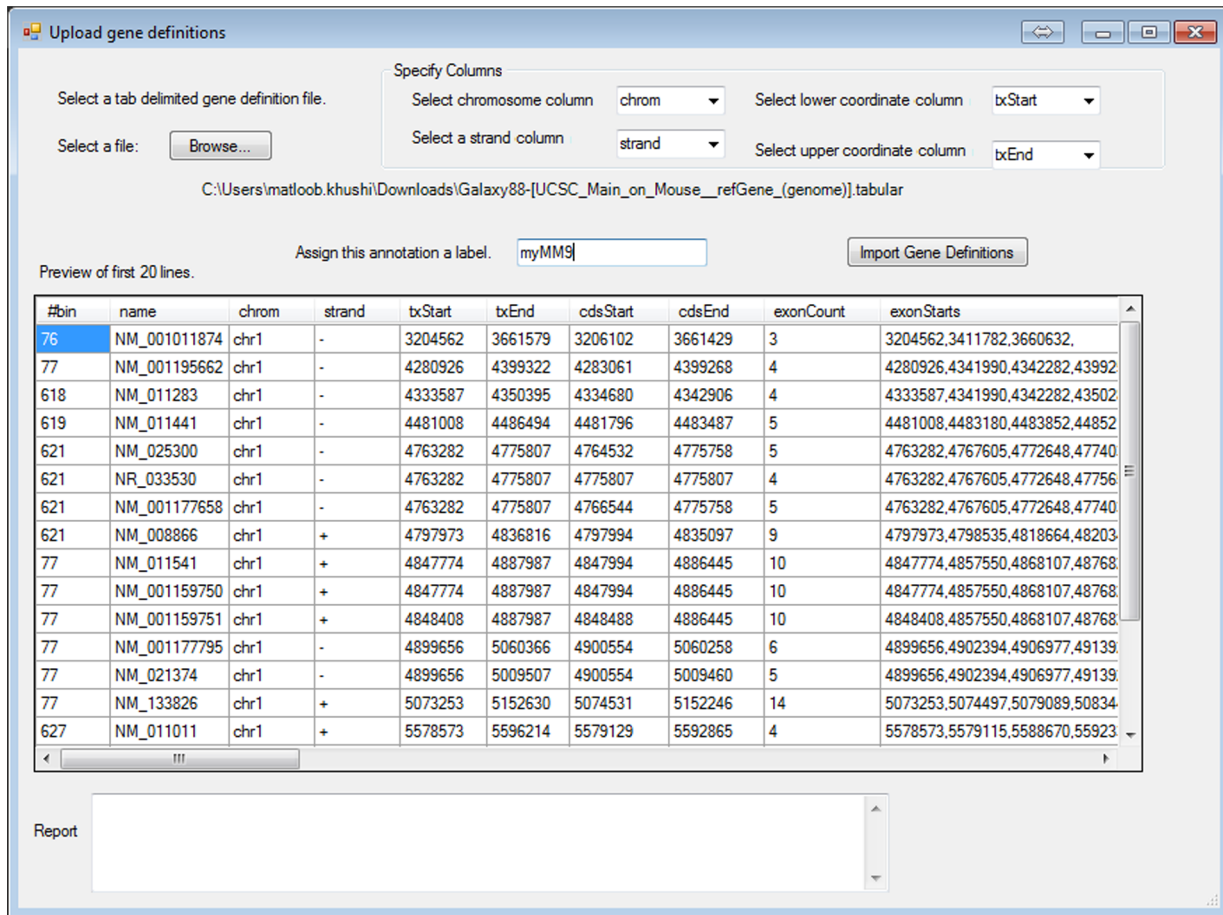


Figure 7. New gene definitions. New gene definitions may be uploaded in the software. The user must specify columns for chromosome, strand, lower and upper coordinates.
doi:10.1371/journal.pone.0087301.g007

uploaded (Figure 7). The delete genes button allows the user to delete the custom uploaded definitions.

vi) Proximal Features. This tab lets the investigator discover features that are in proximity of a gene of interest. The nearby genomic features can be discovered by specifying a locus, chromosome and position (Figure 8A) or a gene (Figure 8B). The gene can be searched by specifying an assembly such as hg19 and typing the exact gene symbol or typing the first few letters of the gene name and pressing the Search button which brings up a list of matching genes. Once a gene is selected, its chromosome, strand, TSS and CDS are displayed and the user can select whether the distance should be calculated from the gene TSS or CDS (Figure 8C). The distance between genomic features and the regions is calculated from the centre of the regions and can be set (Figure 8D). Selecting 'all active datasets' reports cell line, feature/factor and total regions found within the specified distance. If user selects a single KB dataset then full details of all regions within the specified distance are reported which can then be saved back into the KB. All results can also be saved to a file.

vii) Administration. From the Administration tab (Figure 9) users can delete a dataset, save selected data in a tab delimited format, and view regions or region sizes. The distribution of region sizes over the dataset can also be listed or can be visualised as a histogram (Figure 9A) The Clean Up Database button truncates transaction logs, to avoid an impact on software performance.

BiSA Application Example. To present the BiSA utility, we have studied six hg18 datasets available in the KB, transcription factors FoxA1 against FoxA3 [12], CTCF against SA1 [13] and ZNF263 against c-Fos [14]. The forkhead family of pioneer factors (FoxA1, FoxA2 and FoxA3) play important roles in early development to metabolism and homeostasis in adults, and are required for regulation of liver specific genes [12,15,16]. Their DNA-binding domains are highly conserved from yeast to mammals, and there is evidence for cooperative function between the family members [12,16,17]. FOXA factors are pioneer factors due to their ability to bind condensed chromatin and reposition nucleosomes, allowing the binding of other factors [16]. These TFs work together in complex ways to regulate transcription, therefore, the co-location of binding sites of these factors has been extensively studied in the HepG2 cell line [12,18]. Here we demonstrate the application of BiSA by investigating the overlap of binding sites for FoxA1 (8175 regions) and FoxA3 (4598 regions) [12] in HepG2 cells. We have also examined the dataset of Schmidt et al. for the overlap between CTCF and the cohesin complex component SA1 which are known to collocate on DNA [13]. In addition we also studied two non-related transcription factors c-Fos (18211 regions) [14] and ZNF263 (4426 regions) [19] in the K562 (erythromyeloblastoid leukemia) cell line.

BiSA overlap analysis of FoxA1 and FoxA3 with at least 1 bp in common reported 2929 FoxA1 regions. To show this interaction graphically we drew a Venn diagram (Figure 10-A). When we

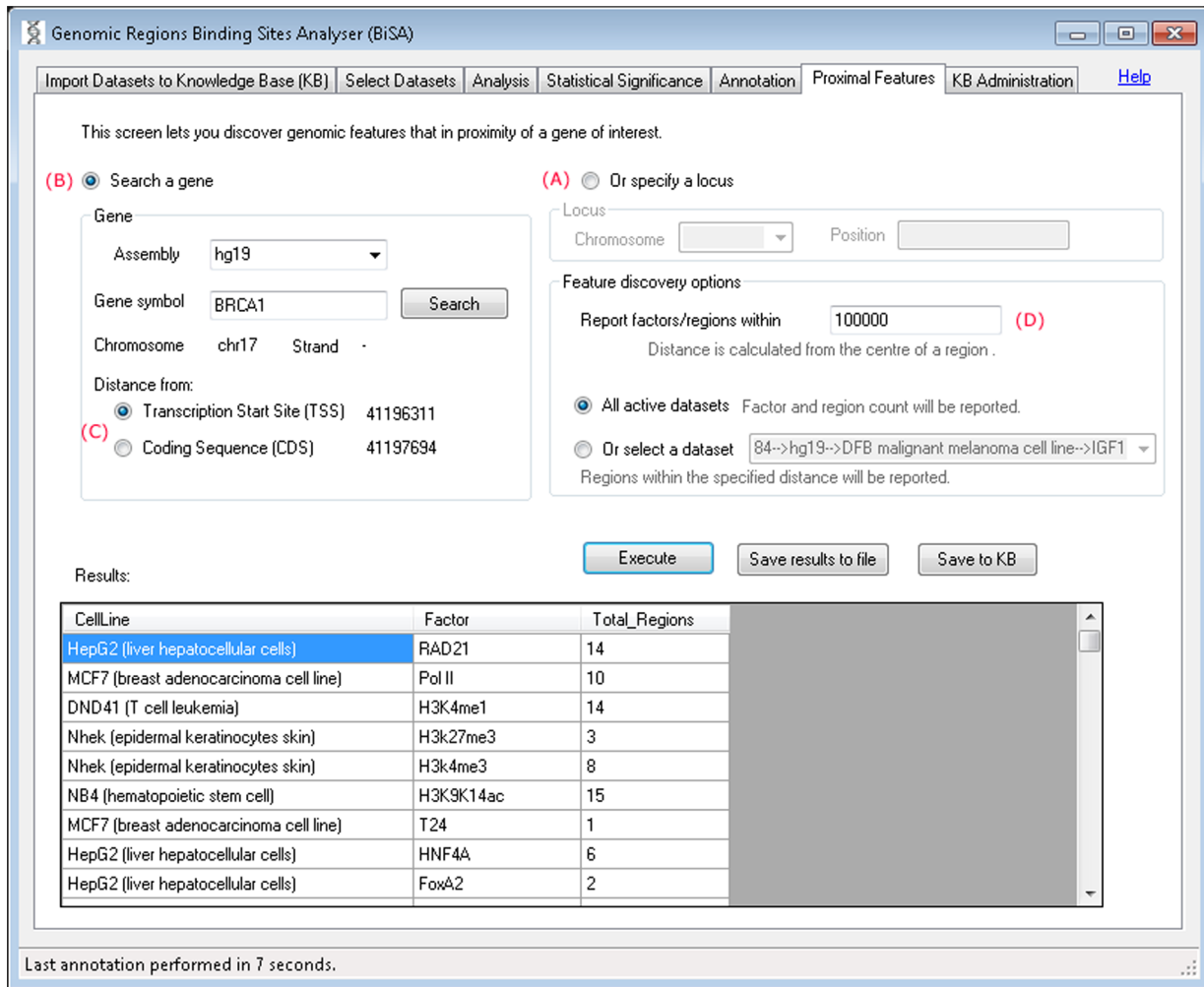


Figure 8. Proximal features. This tab allows users to search for genomic features located in proximity to a specific gene or genomic locus. Searching multiple datasets returns the numbers of binding sites for each factor identified. Selecting a single factor returns detailed binding region information for interactions in proximity to the gene or locus of interest. doi:10.1371/journal.pone.0087301.g008

extracted the overlapping common sections of the regions the number increased to 2939 regions which shows that some regions of the two datasets overlap more than one region of the other dataset. We saved the overlapping sections back into the KB. ‘View Region Sizes’ under the Administration tab is used to draw a histogram of region sizes using bin size 100 (Figure 10-B). The histogram, showing the distribution of overlapping region sizes, reveals that ~99% of overlaps exceed 200 bases and there are more than 1600 regions that have at least 300 bp in common between the two datasets. Similarly the overlap analysis (39,144 common regions) of CTCF (49,243 regions) and SA1 (56,092 regions) is drawn as a Venn diagram and overlapping sections are represented in a histogram (Figure 10-C,D). Similar to the FoxA1-FoxA3 example, the number of common overlapping sections (39586) is greater than the total number of CTCF binding sites due to the fact that a subset of regions overlap multiple regions in the comparison dataset. By contrast, when the unrelated TFs, c-Fos and ZNF263, are compared, just 559 overlaps are detected. A Venn diagram showing the dataset overlap and a histogram summarizing the overlaps are drawn (Figure 10-E,F).

We annotated the common sections of regions to observe their distribution and distance from the nearest TSS by setting criteria

of 100K bp up and downstream from TSS and extracted annotations closest to genes. BiSA reported 3656 gene annotations for FoxA1-FoxA3 overlapping sections, 45,508 annotations for CTCF-SA1 sections and 810 annotations for ZNF263-c-Fos sections. The annotation files also contain the distances from the TSSs.

Finally we investigated the statistical significance of overlap for each of the example comparisons. We calculated p-values for all regions in both datasets for each comparison using the hg18 whole genome domain. Selecting FoxA1 as query and FoxA3 as reference returned an overlap correlation value (OCV) of 0.50. By contrast, if FoxA3 was compared as query to FoxA1 as reference, the OCV was increased to 0.72. This provided an average OCV value of 0.61. An average OCV above 0.5 suggests that two datasets significantly overlap, implying that the overlap between FoxA1 and FoxA3 is statistically significant. The high degree of overlap between CTCF and SA1 also returned a significant average OCV at 0.79. By contrast the lower level of overlap seen between ZNF263 and c-Fos was reflected in a non-significant average OCV of 0.21, confirming that the two TFs are not related and do not act on the same DNA regions in general.

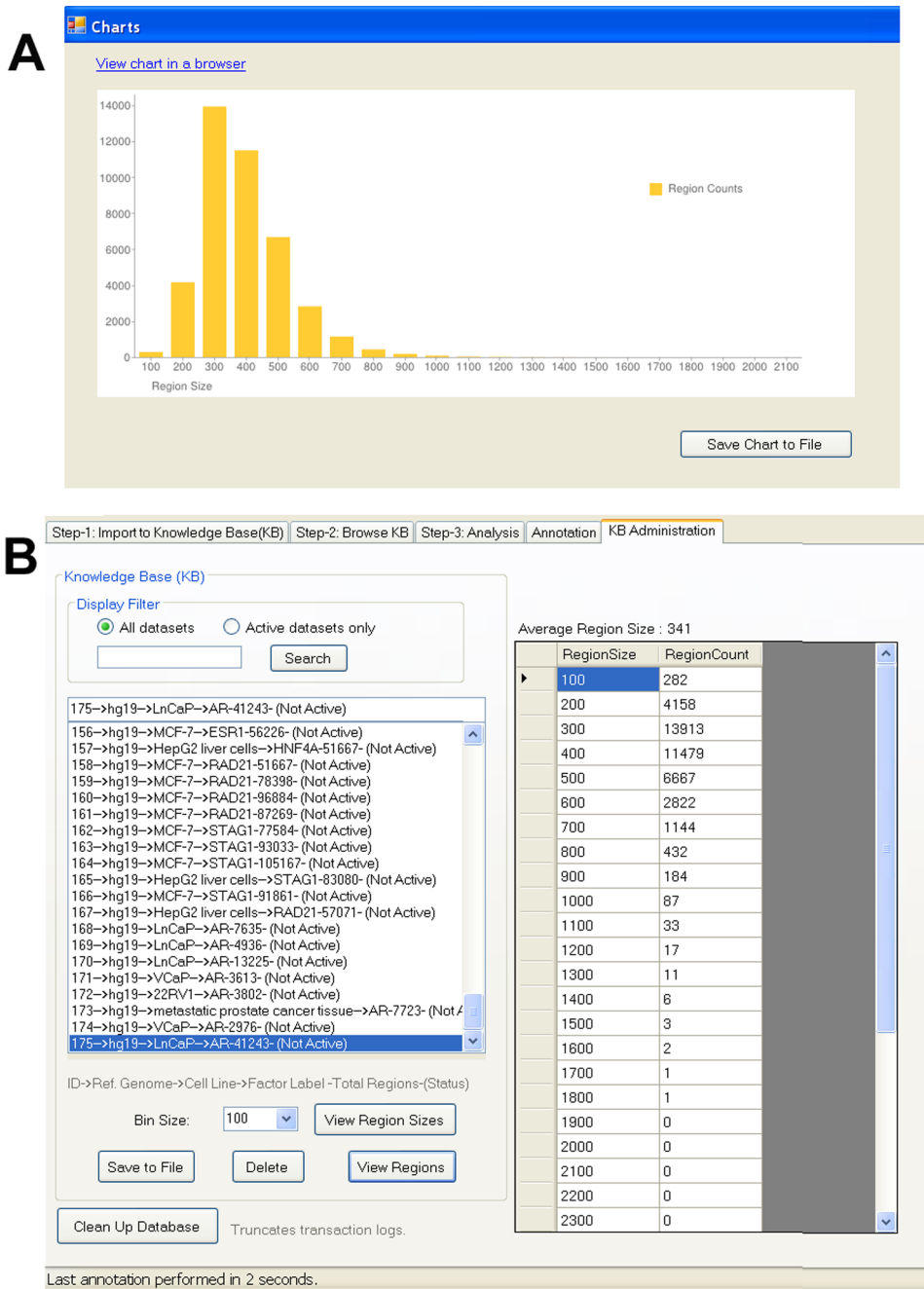


Figure 9. Administration of datasets. From the Administration tab users can delete a dataset, save the data in tab delimited text format, and view the distribution of region sizes over the dataset as a table and histogram.
doi:10.1371/journal.pone.0087301.g009

Discussion

BiSA has been designed to meet the challenges of identifying genomic region overlaps in whole genome datasets. BiSA includes an up-to-date database of previously published studies reporting binding sites for different factors and specific histone modifications in a range of conditions and cell types. No tool, to our knowledge, includes such a pre-loaded knowledge base. Initially we have included data generated from human and mouse cells, and expansion to other organisms is planned. BiSA provides a user-friendly interface allowing the user to define and discover

overlapping and nearby genomic regions either limited by chromosome or genome-wide. Users can visualize genomic overlap results as Venn diagrams and can save chart images for use in publications. BiSA can identify genes associated with binding regions of interest and also the statistical significance of overlapping regions.

Although the Apple Macintosh Unix and Linux environments are popular in genomic research, Windows based informatics tools also exist [20,21,22]. BiSA for Windows exploits the power of today's multi-core personal computers. In comparison to BiSA, most bioinformatics tools are command line, and such tools are

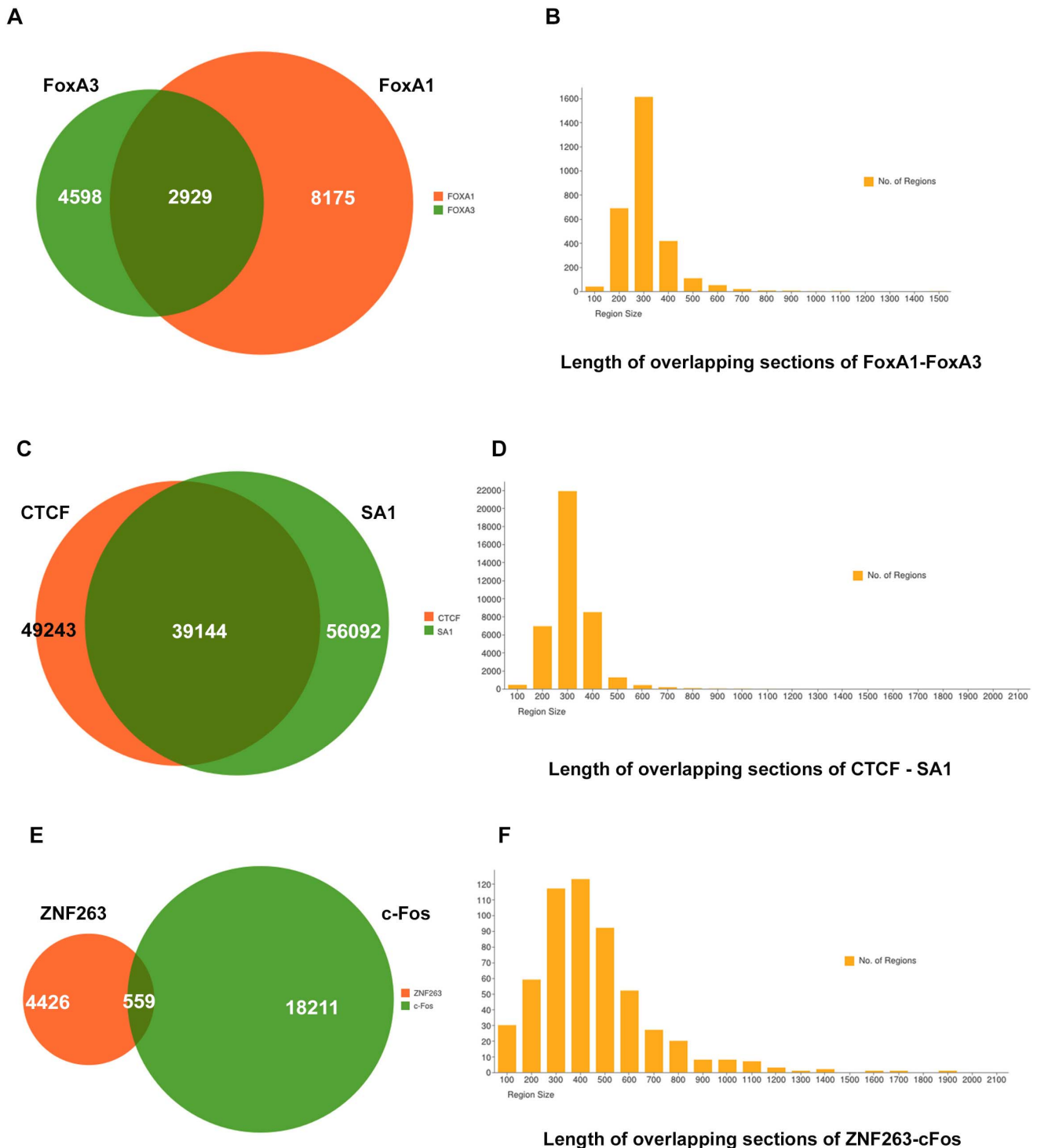


Figure 10. Example study of overlap between FoxA1 and FoxA3, CTCF and SA1, ZNF263 and c-Fos datasets. A) Venn diagram representation of 2,929 overlapping regions in FoxA1 (8,175 regions) and FoxA3 (4,598 regions) datasets. B) Common sections of overlapping regions are saved back into the KB, and for bin size 100 a histogram of the size distribution of region overlaps is drawn. The histogram shows that there are more than 1,600 regions that have at least ~300 bp in common between the two datasets. C) Overlap of 39,144 regions between CTCF and SA1 datasets. D) Distribution of overlapping sections of CTCF and SA1. E) Overlap of 559 regions between ZNF263 and c-Fos datasets. F) Distribution of overlapping sections of ZNF263 and c-Fos. We also observed that in three comparisons >94% of the overlapping sections are >200 bases long, suggesting that overlapping regions usually share a significant section of the two sets.
doi:10.1371/journal.pone.0087301.g010

not easy to install or to operate by the bench biologist. Galaxy [7] offers a web-based tool 'Intersect', however it is limited in functionality. BiSA's Windows GUI is user-friendly for biologists

and provides a sequential step-by-step guide through all the options. BiSA provides an easy interface to search and select KB based on organism, factor, cell line, condition, peak caller or first

author name. We have also developed a BiSA version for Linux/Mac that runs under Galaxy.

A major strength of BiSA is the comprehensive knowledge base, coupled with tools to analyse overlapping regions, statistical significance of the overlapping regions and ability to annotate and visualize the regions of interest. BiSA's comprehensive KB is not only useful for rapid comparison of users' own results to previously published datasets, but also to inform decisions such as selection of a peak caller programme or in comparing numbers of peaks. The KB suggests that MACS is a most popular peak caller software in ChIP-Seq studies followed by Cisgenome and HOMER, whereas, the MAT algorithm is widely used in ChIP-chip studies. In summary, BiSA is designed for ease of use on a Windows platform,

and includes a comprehensive knowledge base of binding site and histone modification datasets. BiSA has the potential to be a useful tool in identifying overlaps in genomic binding regions and histone modifications of common transcription factors.

Acknowledgments

The authors would like to thank Tram Doan for her helpful feedback.

Author Contributions

Conceived and designed the experiments: MK CL JDG CLC. Performed the experiments: MK. Analyzed the data: MK. Wrote the paper: MK. Read, revised and approved the final manuscript: MK CL CLC JDG.

References

- Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, et al. (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature* 489: 91–100.
- Ballaré C, Castellano G, Gaveglia L, Althammer S, González-Vallinas J, et al. (2013) Nucleosome-Driven Transcription Factor Binding and Gene Regulation. *Molecular Cell* 49: 67–79.
- Meyer CA, Tang Q, Liu XS (2012) Minireview: Applications of Next-Generation Sequencing on Studies of Nuclear Receptor Regulation and Function. *Molecular Endocrinology* 26: 1651–1659.
- Heinz S, Benner C, Spann N, Bertolino E, Lin YC, et al. (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* 38: 576–589.
- Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841–842.
- Renaud G, Neves P, Folador EL, Ferreira CG, Passetti F (2011) Segtor: Rapid Annotation of Genomic Coordinates and Single Nucleotide Variations Using Segment Trees. *Plos One* 6: e26715.
- Goecks J, Nekrutenko A, Taylor J, Team TG (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* 11: R86.
- Microsoft (2013) SQL Server Express Edition.
- Project C (2013) Nuclear Receptor Cistrome.
- Chikina M, Troyanskaya O (2012) An effective statistical evaluation of ChIPseq dataset similarity. *Bioinformatics* 28: 607–613.
- MinGW (2013) MinGW | Minimalist GNU for Windows.
- Motallebipour M, Ameer A, Reddy Bysani MS, Patra K, Wallerman O, et al. (2009) Differential binding and co-binding pattern of FOXA1 and FOXA3 and their relation to H3K4me3 in HepG2 cells revealed by ChIP-seq. *Genome Biology* 10: R129.
- Schmidt D, Schwalie PC, Ross-Innes CS, Hurtado A, Brown GD, et al. (2010) A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res* 20: 578–588.
- Raha D, Wang Z, Moqtaderi Z, Wu L, Zhong G, et al. (2010) Close association of RNA polymerase II and many transcription factors with Pol III genes. *Proc Natl Acad Sci U S A* 107: 3639–3644.
- Lee CS, Friedman JR, Fulmer JT, Kaestner KH (2005) The initiation of liver development is dependent on Foxa transcription factors. *Nature* 435: 944–947.
- Friedman JR, Kaestner KH (2006) The Foxa family of transcription factors in development and metabolism. *Cell Mol Life Sci* 63: 2317–2328.
- Soccio RE, Tuteja G, Everett LJ, Li Z, Lazar MA, et al. (2011) Species-specific strategies underlying conserved functions of metabolic transcription factors. *Mol Endocrinol* 25: 694–706.
- Wallerman O, Motallebipour M, Enroth S, Patra K, Bysani MS, et al. (2009) Molecular interactions between HNF4a, FOXA2 and GABP identified at regulatory DNA elements through ChIP-sequencing. *Nucleic Acids Res* 37: 7498–7508.
- Frietze S, Lan X, Jin VX, Farnham PJ (2010) Genomic targets of the KRAB and SCAN domain-containing zinc finger protein 263. *J Biol Chem* 285: 1393–1403.
- Ji H, Jiang H, Ma W, Johnson DS, Myers RM, et al. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotech* 26: 1293–1300.
- Khushi M, Carpenter J, Balleine R, Clarke C (2012) Development of a data entry auditing protocol and quality assurance for a tissue bank database. *Cell and Tissue Banking* 13: 9–13.
- Khushi M, Carpenter J, Balleine R, Clarke C (2012) Electronic Biorepository Application System: Web-Based Software to Manage Receipt, Peer Review, and Approval of Researcher Applications to a Biobank Biopreservation and Biobanking 10: 37–44.

Bioinformatic analysis of cis-regulatory interactions between progesterone and estrogen receptors in breast cancer

Matloob Khushi*, Christine L. Clarke and J. Dinny Graham

Centre for Cancer Research, Westmead Millennium Institute, Sydney Medical School—Westmead, University of Sydney, Australia

* Current affiliation: Bioinformatics Unit, Children's Medical Research Institute, Westmead, NSW, Australia

ABSTRACT

Chromatin factors interact with each other in a cell and sequence-specific manner in order to regulate transcription and a wealth of publically available datasets exists describing the genomic locations of these interactions. Our recently published BiSA (Binding Sites Analyser) database contains transcription factor binding locations and epigenetic modifications collected from published studies and provides tools to analyse stored and imported data. Using BiSA we investigated the overlapping cis-regulatory role of estrogen receptor alpha ($ER\alpha$) and progesterone receptor (PR) in the T-47D breast cancer cell line. We found that $ER\alpha$ binding sites overlap with a subset of PR binding sites. To investigate further, we re-analysed raw data to remove any biases introduced by the use of distinct tools in the original publications. We identified 22,152 PR and 18,560 $ER\alpha$ binding sites (<5% false discovery rate) with 4,358 overlapping regions among the two datasets. BiSA statistical analysis revealed a non-significant overall overlap correlation between the two factors, suggesting that $ER\alpha$ and PR are not partner factors and do not require each other for binding to occur. However, Monte Carlo simulation by Binary Interval Search (BITS), Relevant Distance, Absolute Distance, Jaccard and Projection tests by Genometricorr revealed a statistically significant spatial correlation of binding regions on chromosome between the two factors. Motif analysis revealed that the shared binding regions were enriched with binding motifs for $ER\alpha$, PR and a number of other transcription and pioneer factors. Some of these factors are known to co-locate with $ER\alpha$ and PR binding. Therefore spatially close proximity of $ER\alpha$ binding sites with PR binding sites suggests that $ER\alpha$ and PR, in general function independently at the molecular level, but that their activities converge on a specific subset of transcriptional targets.

Submitted 21 May 2014
Accepted 15 October 2014
Published 18 November 2014

Corresponding author
Matloob Khushi,
mkhushi@uni.sydney.edu.au

Academic editor
Kenta Nakai

Additional Information and
Declarations can be found on
page 14

DOI 10.7717/peerj.654

© Copyright
2014 Khushi et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

Subjects Bioinformatics, Computational Biology, Molecular Biology

Keywords Transcription factors, Estrogen receptor alpha, Progesterone receptor, $ER\alpha$, ESR1, PR, Breast cancer, T47D, BiSA, Genomic region database

INTRODUCTION

The ovarian steroid hormones progesterone and estrogen play critical roles in the development and progression of breast cancer and endometriosis (*D'Abreo & Hindenburg, 2013; Salehnia & Zavareh, 2013; Shao et al., 2014*). These hormones exert their functions

by activating specific nuclear receptors, estrogen binds to estrogen receptor (ER α) and progesterone binds to progesterone receptor (PR) (Tsai & O'Malley, 1994).

Once activated these receptors bind to their DNA response elements and regulate transcription of target genes. ER α and PR, along with human epidermal growth factor receptor 2 (HER2), are used to classify phenotypes in breast cancers and to predict response to specific therapies (Cadoo, Fornier & Morris, 2013; Kittler et al., 2013). A high number of ER α positive breast cancers are also PR positive (Cadoo, Fornier & Morris, 2013; Penault-Llorca & Viale, 2012). Furthermore, studies from animal models and clinical trials have shown that progesterone via its receptor PR is a major player in development and growth of breast cancer and uterine fibroids, however, PR inhibits the development of estrogen-driven endometrial cancer (Ishikawa et al., 2010; Kim, Kurita & Bulun, 2013). Many recent reviews highlight the importance of the role that progesterone and estrogen play via their receptors in various types of breast cancers (Abdel-Hafiz & Horwitz, 2014; Kalkman, Barentsz & van Diest, 2014; Obiorah et al., 2014; Wang & Di, 2014; Yadav et al., 2014). Therefore it is important to understand how ER α and PR work together in regulating a number of cellular pathways, and clinical and molecular research on these factors continue to unveil new insights (Bulun, 2014).

It is acknowledged that ER α and PR binding, as well as that of other steroid hormone receptors, is assisted by binding of the pioneer transcription factor FOXA1 (Ballare et al., 2013; Lam et al., 2013) to condensed chromatin, therefore, the interactions of FOXA1 with other factors have been well studied (Augello, Hickey & Knudsen, 2011; Bernardo & Keri, 2012). There are a number of publications that have studied PR binding sites in progesterone-treated breast and other tissues (Ballare et al., 2013; Clarke & Graham, 2012; Yin et al., 2012). Many studies have also published ER α binding sites (Joseph et al., 2010; Schmidt et al., 2010; Tsai et al., 2010). However there is lack of investigation into the combined action of the two factors on DNA. Therefore in this report we investigated the interaction of these nuclear receptors on DNA. Our previously published BiSA database (Khushi et al., 2014) contains a number of datasets describing ER α and PR binding sites for various cell lines, therefore, we investigated the binding pattern of these factors in the T-47D breast cancer cell line. T-47D cells are derived from metastatic female human breast cancer and are known to be ER α and PR positive and their growth is simulated by the treatment of estrogen (Chalbos et al., 1982; Ström et al., 2004).

METHODS

PR data were taken from the study of Clarke & Graham (2012) and ER α data were obtained from the ENCODE project (Gertz et al., 2012). PR data were obtained by treating T47D cells with the progestin ORG2058 for 45 min, followed by PR-specific chromatin immunoprecipitation and deep sequencing (ChIP-Seq). Gertz et al. studied ER α binding sites by treating with estradiol (E2), GEN (Genistein) and BPA (Bisphenol A) and conclude that compared to E2, GEN and BPA treatment results in fewer ER α binding sites and less change in gene expression. We selected the E2-treated dataset for our study. Datasets from both studies were of 36 base pair lengths on the Illumina platform. The PR data were




generated using an Illumina Genome Analyzer Iix while ER α libraries were sequenced on Illumina HiSeq 2000. The data used in this study have been derived from peer-reviewed publications, suggesting that they are of an acceptable quality, in addition we also ensured standard quality control checks prior to our re-analysis of the raw data. The two studies used different genome assemblies and different tools to align the reads and to call the peaks. Therefore, to remove any biases we re-analysed the raw ER α and PR data. We mapped the raw data to the GRCh37/hg19 assembly using Bowtie version 2 (Langmead & Salzberg, 2012). The aligned replicates were merged using Picard tools (Li et al., 2009) and Model-based Analysis of ChIP-seq Algorithm (MACS) version 1.4.2 (Zhang et al., 2008) was employed, with default settings, to identify PR and ER α binding regions in the two datasets. Regions associated with greater than 5% false discovery rate (FDR) were removed (Zhang et al., 2008).

We performed motif analysis using HOMER software (Heinz et al., 2010). HOMER employs a differential motif discovery algorithm by comparing two sets of sequences and quantifying consensus motifs that are differentially enriched in a set. HOMER automatically generates an appropriate background sequence matched for the GC content to avoid bias from CpG Islands. The tool is exclusively written for analysing DNA regulatory elements in ChIP-Seq experiments and has been used in number of high impact publications (Berman et al., 2012; Wang et al., 2011b; Xie et al., 2013).

Overlapping features were studied in BiSA (Khushi et al., 2014). BiSA is a bioinformatics database resource that can be run on Windows as a personal resource or web-based under Galaxy (Goecks et al., 2010) as a collaborative tool. BiSA is pre-populated with published transcription factor and histone modification datasets and allows investigators to run a number of overlapping and non-overlapping genomic region analyses using their own datasets, or against the pre-loaded Knowledge Base. Overlapping features can be visualised as a Venn diagram and binding regions of interest can also be annotated with nearby genes. BiSA also provides an easy graphical interface to find the statistical significance of observed overlap between two genomic region datasets by implementing the IntervalStat tool (Chikina & Troyanskaya, 2012). The tool calculates a p -value for each peak region by comparing a region from the query dataset to all regions in a reference dataset. The tool restricts the analysis to regions that are within a domain dataset which can be a whole genome or can be possible interval locations such as promoter proximal regions. Based on IntervalStat calculated p -values BiSA calculates a summary statistic that we refer to as the Overlap Correlation Value (OCV). The OCV ranges from 0 to 1, the closer the value to 1 the stronger the significance of overlap of two datasets. The OCV represents the fraction of regions in the query dataset with a p -value less than a specified threshold. In BiSA, we have set the threshold p -value to 0.05 and used a number of domains such as whole genome and promoter proximal regions for this analysis.

We also investigated the spatial correlation of regions of whole datasets being closer to each other by Binary Interval Search (BITS) (Layer et al., 2013) and Genometricorr (Favorov et al., 2012). BITS implements a Monte Carlo simulation by comparing actual overlapping regions to random observed overlap. Genometricorr considers one genomic

Table 1 Motif analysis of PR regions. Known motif analysis of PR top 1,000 regions using Homer software.

Motif	Name	P-value	% of targets sequences with motif
	PR(NR)/T47D	1e-123	59.40%
	FOXA1(Forkhead)/LNCAP-FOXA1	1e-28	27.10%
	AP-2gamma(AP2)/MCF7-TFAP2C	1e-10	13.70%




region set as a reference and other set as a query and provides four asymmetric pair-wise statistical tests (i) relative distance also called local correlation, (ii) absolute distance, (iii) Jaccard statistic and (iv) projection statistical tests. In local correlation the significance of relative distance between the genomic regions is measured by Kolmogorov–Smirnov test, in absolute distance test the significance of base pair distance among the regions is measured by permutation test, Jaccard statistic takes into account the ratio of intersecting bases to the union base pairs. A projection test calculates the overlapping centre points of query to reference regions and finds the significance of result outside of the null expectation by binomial test (*Favorov et al., 2012*). We performed 10,000 simulations for BITS and Genometricorr statistical tests.

We performed functional annotation of ER α -PR common cis-regulatory regions using GREAT (Genomic Regions Enrichment of Annotations Tool) (*McLean et al., 2010*). GREAT incorporates annotations from 20 ontologies covering gene ontology, phenotype data, human disease pathways, gene expression, regulatory motifs and gene families. We performed GREAT annotation using its default settings. A region was considered to have a proximal association with a gene if it was within 5 kb upstream or 1 kb downstream of the transcription start site (TSS). Regions outside this distance and up to 1,000 kb from the TSS to the next gene proximal region were considered to have a distal association.

RESULTS

Analysis of PR and ER α ChIP-seq data from T-47D breast cancer cells revealed 22,152 PR and 18,560 ER α binding regions with FDR <5%. HOMER motif analysis on the top ranked 1,000 regions by peak score revealed the strong presence of a PRE motif (59.40%) and ERE motif (48.80%) (*Tables 1 and 2*). These were the most statistically significant motifs identified, in agreement with other studies (*Kim, Kurita & Bulun, 2013; Lin et al., 2007*). In addition, in PR binding regions we found motifs for the transcriptional partners FOXA1 and AP-2 (TFAP2C) as other top ranked motifs. The transcription factor activator protein 2C (TFAP2C) is known to be involved in normal mammary development, differentiation, and oncogenesis (*Cyr et al., in press; Lal et al., 2013; Woodfield et al., 2010*). Interestingly PR motifs were present in 344 (34.4%) of the 1,000 top ranked ER α binding regions. Consensus FOXA1 motifs were also detected in 27% of PR binding regions and 24% of regions bound by ER α . FOXA1 is a member of the forkhead family of transcription factors, which are known to bind and reconfigure condensed chromatin to

Table 2 Motif analysis of ER α regions. Known motif analysis of ESR1 top 1,000 regions.

Motif	Name	P-value	% of targets sequences with motif
	ERE(NR/IR3)/MCF7-ERa	1e-474	48.80%
	FOXA1(Forkhead)/LNCAP-FOXA1	1e-22	24.30%
	PR(NR)/T47D-PR	1e-20	34.40%

enable the binding of other transcription factors (*Bernardo & Keri, 2012*). The presence of high quality (p -value $< 1.00e-05$) peaks and known conserved PR and ER α recognition sequences confirmed the success of the alignment and peak-calling process.

The size distribution of ER α (18,560 regions) and PR (22,152 regions) binding regions were visualised by drawing a histogram and box plot (*Figs. 1 and 2*). Mean PR binding region size was 1508 with a median of 1336. In contrast, ER α binding regions were on average half the size of PR binding regions, with a mean size of 601 and median 529. Most PR binding regions ($\sim 94\%$) were greater than 1 kb, whereas most ER α binding regions ($\sim 95\%$) were less than 1 kb. The longer PR regions may be due to longer input DNA fragment lengths in the original samples (*Kharchenko, Tolstorukov & Park, 2008; Landt et al., 2012*).

Limited overlap of ER α and PR regions

Using BiSA, we identified that almost one quarter (23.6%) of ER α binding regions (4,344) overlap with 3,870 unique PR binding regions. This revealed that some long PR binding regions spanned more than one ER α binding region and the reverse was also true for large ER α binding regions. In total, we found 4,358 sections that were common to the two datasets. The Venn diagram in *Fig. 3A* shows this overlap between the two ligand-activated transcription factors. The 4,358 overlapping sections of the regions common to the two datasets were extracted and plotted for their region lengths (*Fig. 3B*). Out of 4,358 overlapping sections 4,279 (98.2%) were more than 100 bases long, suggesting a strong binding overlap between the two transcription factor data sets. An example of a shared ER α and PR binding region is shown in *Fig. 4*. The 631 bp ER α binding region (red dotted lines) is completely contained within the 813 bp PR binding region (blue dotted lines) and the two regions share the peak centre location (*Fig. 4*).

Statistical analysis of ER α -PR overlap

To determine whether the overlap between ER α and PR binding was statistically significant, statistical analysis was performed in BiSA, BITS and Genometricorr. In BiSA, using a whole genome domain and selecting the ER α cistrome as query and PR as reference revealed an overlap correlation value of 0.33. The value decreased to 0.26 when PR was selected as query and ER α as reference. This showed that, although a considerable proportion of ER α binding regions are also bound by PR, the two receptors

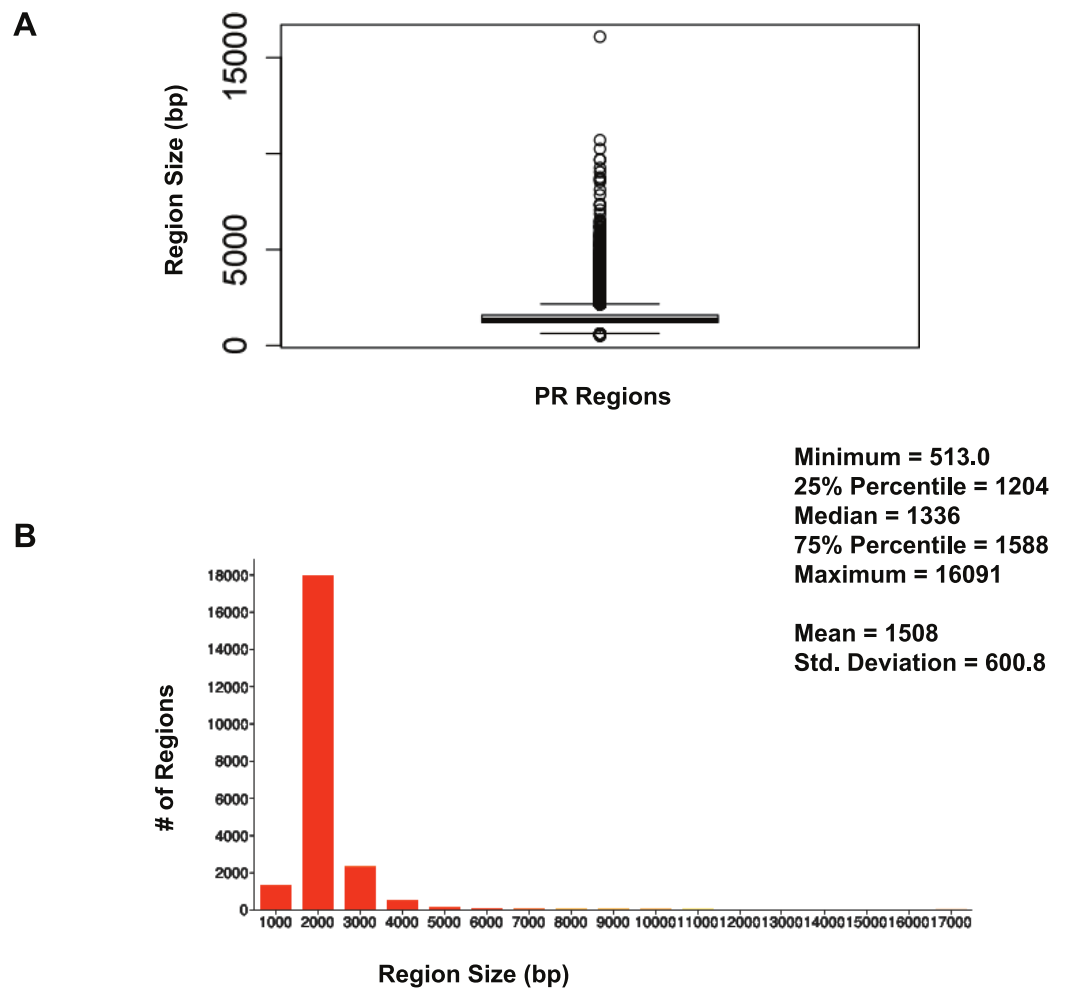


Figure 1 Distribution of PR binding region sizes. (A) Box plot with mean and median information. (B) Histogram of region sizes with bin size 1,000.

do not cooperate for binding at all sites. To determine whether the significance of ER α -PR binding overlap was greater in functionally relevant genomic regions, we compared the level of binding overlap over a range of genomic domains from promoter proximal (within 500 b of a TSS) to more distal regions (Table 3). We found a low though consistent overlap correlation value (~ 0.3) whether promoter proximal or distal sites were included in the analysis (Table 3). To confirm that the OCV result is independent of the mean region sizes of the two datasets, we fixed the PR region sizes to 300 bases from each side of peak summits to match mean ER α region length (mean = 601) and performed the OCV test again. This did not change the OCV (0.33) for the whole genome dataset, and there was negligible change in OCV observed for other domains (Table 3).

Using BITS and Genometricorr, we further investigated whether the spatial proximity correlation between PR and ER α binding was more significant than expected by chance. BITS Monte Carlo simulation reported that the spatial correlation of ER α and PR was statistically significant, with a p -value of 0.0001. Similarly Genometricorr's Relative

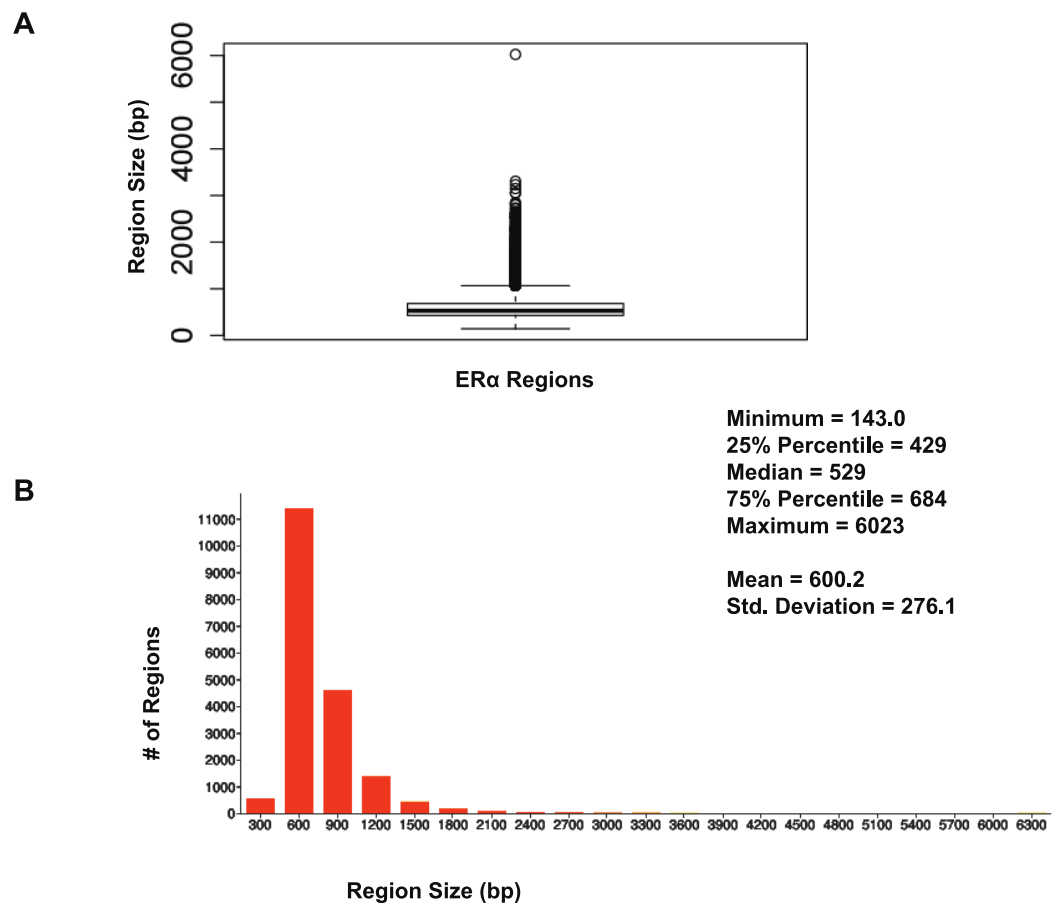


Figure 2 Distribution of ER α binding region sizes. (A) Box plot with mean and median information. (B) Histogram of ER α region sizes with bin 200.

Correlation test, Absolute Distance test, Jaccard test and Projection tests also reported the spatial correlation between the two factors as statistically significant (p -value = $< 1e-04$) (Fig. 5). We repeated the tests for the 600bp fixed-width PR dataset and found no change in reported p -values from BITS or Genometricorr. This confirmed that a change in average region size between the two datasets does not affect the statistical analysis and demonstrated that the tendency for binding events for the two factors to be close to each other is statistically significant. Therefore we conclude that, although there are a number of statistically significant shared binding sites in the ER α and PR datasets, and that ER α and PR often bind in proximity to each other, the observed overlap of the two factors is not strong enough for them to be considered as co-factors that consistently co-operate on shared binding regions. However, the close proximity of the binding regions for the two factors shows a spatial convergence and is statistically significant.

Motif analysis

The 4,358 common sections of ER α -PR were searched for known motifs. Known motif analysis in these common sections revealed a strong presence of ERE, forkhead protein

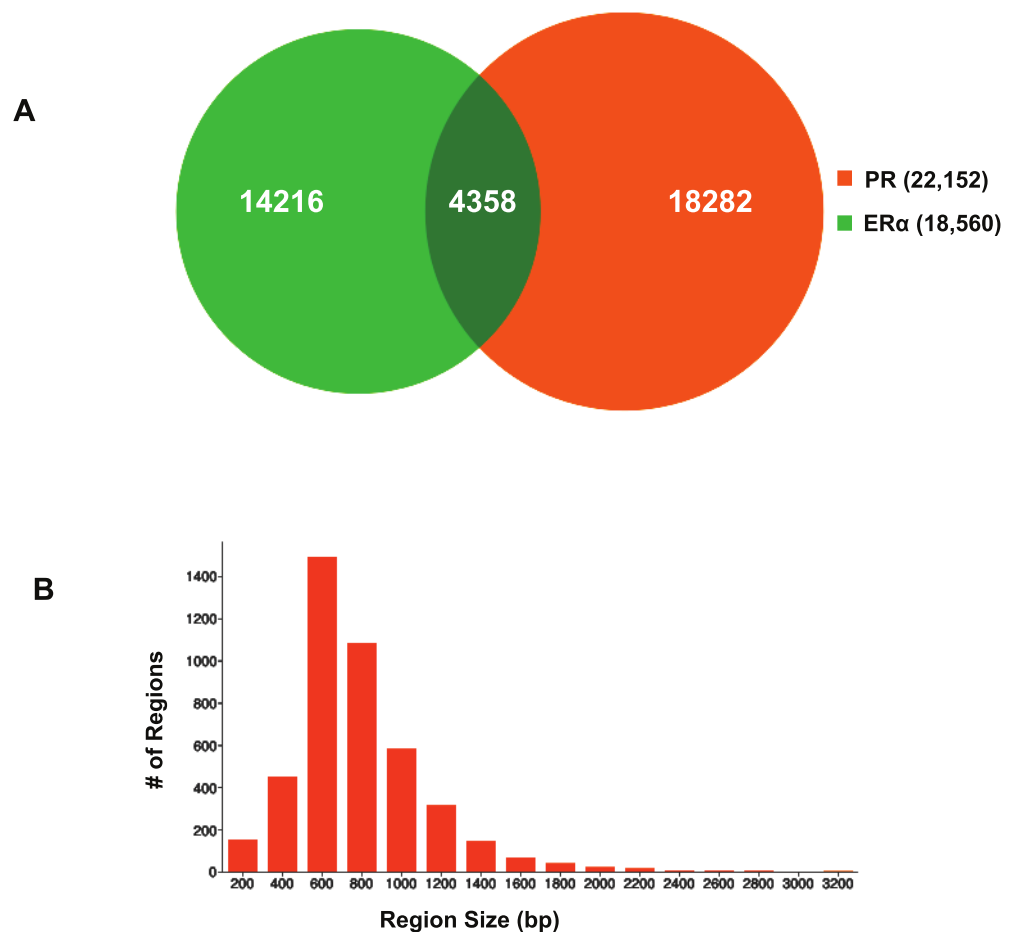


Figure 3 Visualisation of ER α and PR binding region overlap. (A) Venn diagram showing overlap between ER α and PR data. The 4,344 ER α binding regions overlap with 3,870 unique PR binding regions making up 4,358 overlapping sections. (B) Region sizes of 4,358 regions common to the ER α and PR datasets.

and PRE motifs. In [Table 4](#), we listed the top ranked motifs, ordered by p -value. A PRE motif was found in 41.88% (1,825) of the total 4,358 regions, which was much higher than the number of ERE motifs detected 14.3% (623) of the sequences. However, this may reflect the higher stringency of the position specific scoring matrix used to identify ERE motif occurrence than the matrix used to find PRE motifs since the p -value for ERE motif detection ($1e-291$) was much stronger than the p -value for PRE motif occurrence in the dataset ($1e-179$). The presence of FOXA1 motifs in these regions confirms that the factor facilitates the binding of ER α and PR on these regions as previously reported ([Augello, Hickey & Knudsen, 2011](#); [Bernardo & Keri, 2012](#); [Nakshatri & Badve, 2009](#)). In addition AP-2 and TEAD4 (TEA) motifs were also identified in these regions and in the 1,000 top scoring PR binding regions. AP-2 has a known role in normal mammary development and breast cancer ([Cyr et al., in press](#); [Lal et al., 2013](#); [Woodfield et al., 2010](#)). TEAD4 has also been shown to be co-expressed with other oncogenes and is correlated with poor prognosis ([Xia et al., 2014](#); [Mesrouze et al., 2014](#); [Lim et al., 2014](#)). The presence of the related motifs in

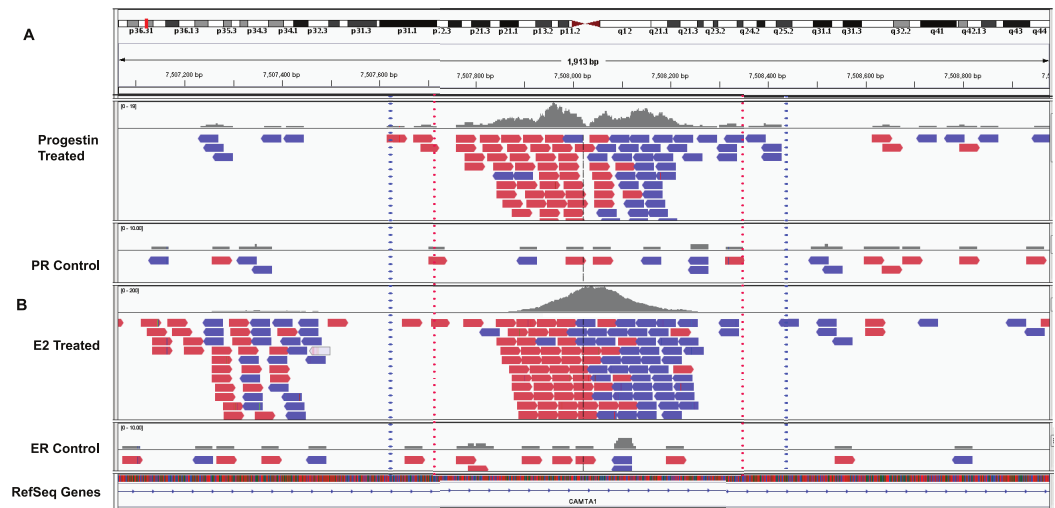


Figure 4 Example overlapping region. IGV snapshot of PR binding region at chr1:7507615–7508428 (marked by blue dotted lines) and ER α binding region (marked by red dotted lines). (A) Progestin treated and control samples. (B) Estradiol (E2) treated and control sample. The red boxes are reads that mapped to the forward strand and blue boxes are reads that mapped to the reverse strand of the human genome (build hg19).

Table 3 BiSA Overlap Correlation Value (OCV) testing. BiSA Statistical analysis of overlap between ER α and PR datasets using different domain datasets.

Domain	Overlap Correlation Value (OCV)			# of overlaps ^b /total ER α regions in domain
	Query = ER α Reference = PR	Query = PR Reference = ER α	Query = ER α Reference = PR (600 bp long) ^a	
Whole Genome	0.33	0.26	0.33	4,344/18,560
500 bp upstream, downstream of TSS	0.3	0.17	0.22	112/419
1 kb upstream, downstream of TSS	0.28	0.18	0.25	157/647
5 kb upstream of TSS	0.3	0.21	0.28	304/1,224
5 kb upstream, downstream of TSS	0.31	0.22	0.3	522/2,147
10 kb upstream, downstream of TSS	0.31	0.22	0.3	929/3,666
45 kb–55 kb upstream of TSS	0.29	0.21	0.28	449/1,929
95 kb–105 kb upstream of TSS	0.31	0.24	0.3	514/2,017
90 kb–110 kb upstream of TSS	0.31	0.23	0.3	878/3,495

Notes.

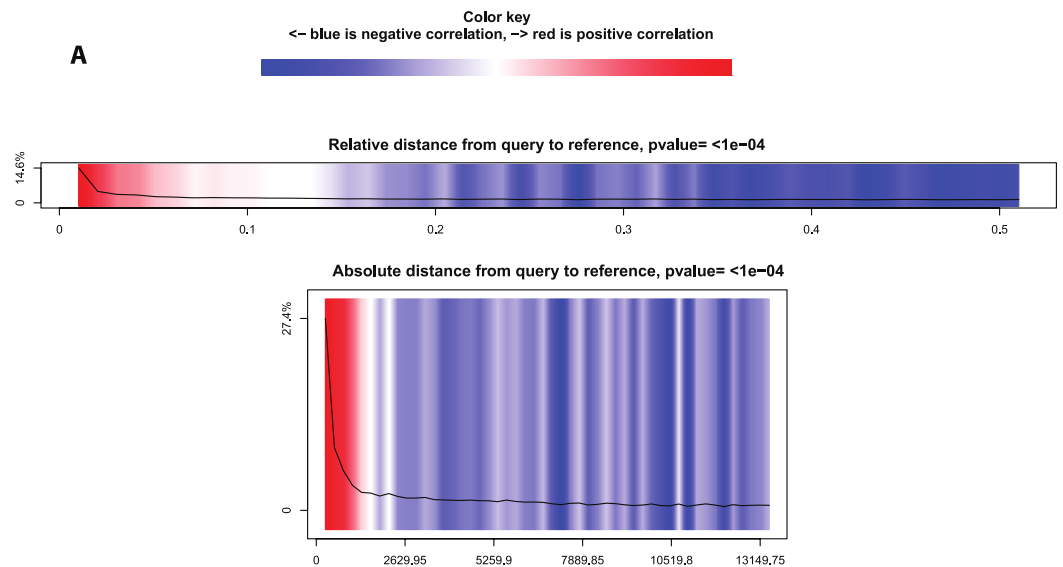
^a PR regions are fixed to 600 bp long by cutting off 300 bp on both sides of peak summits.

^b Number of overlaps in this column is reported by selecting ER α as the query and PR as the reference dataset.

the ER α -PR shared regions as well as in regions that bind uniquely ER α or PR suggests that AP-2 and/or TEAD play a key role for both receptors and could be important in facilitating cooperation between the two nuclear receptors.

Using Homer, we also looked at relative position distributions of these motifs (Fig. 6). We found that the motifs converge around the centres of the peaks, supporting their biological significance as primary binding events.

Overlay line on graph is data density, over 50 bins
 This range of densities is real though does not on its own convey significance
 The p-value signals whether the trends are statistically significant.



B

Results: All chromosomes

Overlap summary (Jaccard and projection tests)

Jaccard p-value: $<1e-04$

Query and reference intervals overlap significantly more than expected by chance, by Jaccard

Query midpoints and reference intervals overlap significantly more than expected by chance, by projection

Figure 5 Statistical significance test using Genometricorr. Genometricorr statistical significance analysis of ER α (query)-PR (reference). (A) Relative and Absolute Distance Correlation tests are shown graphically. Overlay line (data density) when in the blue section shows negative correlation while the high density in the red section shows positive correlation. (B) Results from Jaccard and Projection tests are shown in text.

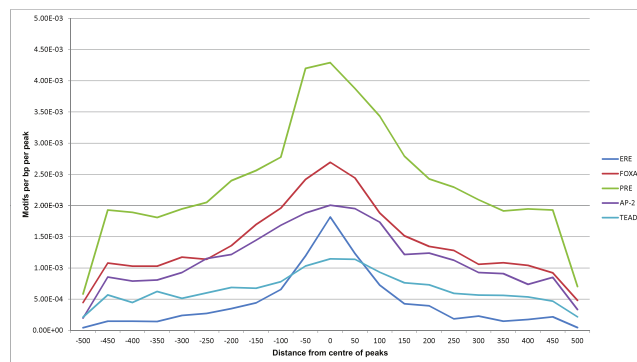







Figure 6 Motif position distributions in ER α -PR overlapping regions. Frequency distribution of ERE, FOXA1, PRE, AP-2 and TEAD4 motifs around centres of peaks using a 50 bp bin size.

Table 4 Known motif analysis of ER α and PR overlapping common regions. Top ranked known motif analysis of ER α -PR common sections (4,358 regions).

Motif	Name	P-value	% of targets sequences with motif
	ERE(NR/IR3)/MCF7-ERa	1e-291	14.30%
	FOXA1(Forkhead)/LNCAP-FOXA1	1e-249	35.11%
	PR(NR)/T47D-PR	1e-179	41.88%
	AP-2gamma(AP2)/MCF7-TFAP2C	1e-122	20.38%
	TEAD4(TEA)/Tropoblast-Tead4	1e-86	17.97%

Enrichment analysis of ER α -PR common regions

We used GREAT (Genomic Regions Enrichment of Annotations Tool) ([McLean et al., 2010](#)) to interpret the functional role of 4,358 ER α -PR common regions. GREAT revealed that only 34 regions ($\sim 0.8\%$) are not associated with any gene and 3,687 ($\sim 85\%$) regions are associated with 2 genes ([Fig. 7](#)). Most of the regions were found to be distal binding events while 405 ($\sim 9\%$) regions are within 5 kb of transcription start sites (TSS). Region to gene association revealed MYC has the maximum number of regions linked to this gene (26 regions). The known role of estrogen-induced MYC oncogene in breast cancer ([Orr et al., 2012](#); [Wang et al., 2011a](#)) confirms a biological relevant regions-to-gene association. PGR was also among the top 10 genes identified with the largest number of associated regions ([File S1](#)). Gene ontology enrichment analysis of the common regions revealed epithelial cell development as the most significant biological process ([File S1](#)). Epithelial cell development was linked to 30 genes associated with 120 regions out of which 4 regions were within 5 kb of a TSS. Pathway Commons, a meta-database of public biological pathway information ([Cerami et al., 2006](#)), revealed the ER α signalling network as the most significant term (p -value = $5.7e-37$) where 137 regions were found regulating 24 genes associated with this pathway. The FOXA1 transcription factor network and IL6-dedicated signalling events were also significant terms (p -value $1.6e-19$ and $2.6e-17$). Mouse phenotype analysis revealed two breast cancer related ontologies (abnormal mammary gland epithelium physiology and abnormal mammary gland development) as the most significant terms. There were 32 regions associated with 5 genes linked to abnormal mammary gland epithelium physiology and 189 regions associated with 52 genes linked to mammary gland development. The [File S1](#) also lists regions and associated genes with the ontologies.

DISCUSSION

The BiSA database provides a good starting point for studying overlapping binding by a range of transcription factors from a comprehensive collection of published studies ([Khushi et al., 2014](#)). The datasets available in BiSA represent the original genomic

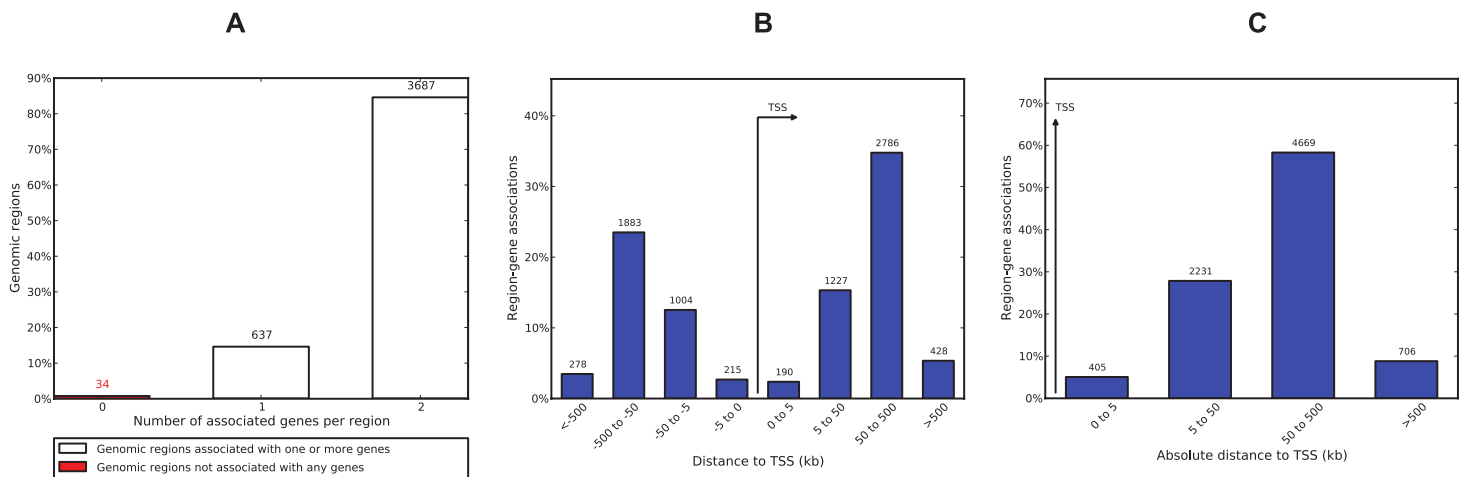


Figure 7 ER α -PR common region-gene association. (A) Number of associated genes per region. (B) Region-gene association binned by orientation and distance to TSS. (C) Region-gene association binned by absolute distance to TSS.

locations identified in the published studies from which they are sourced. Although the same standard pipeline has often been applied, it must be acknowledged that differences in read alignment algorithms (Kerpedjiev et al., 2014; Lunter & Goodson, 2011) and the use of a variety of peak-caller programmes (Ladunga, 2010; Pepke, Wold & Mortazavi, 2009; Wilbanks & Facciotti, 2010) has an impact on downstream analysis, largely due to differences in stringency that affects the number of genomic regions identified. Our initial investigation of the overlap in ER α and PR binding in T-47D cells, utilizing the published binding regions, revealed an overlap of $\sim 27\%$ of ER α binding regions with the published PR cisome (data not shown). This suggested an interesting functional relationship between the receptors, which justified further study. To perform a more rigorous exploration of their overlapping binding patterns, we reanalysed the raw ER α and PR ChIP-seq data using a standardized pipeline. This illustrates the great value of BiSA as an easy to implement first pass tool to investigate potential functional relationships in transcription factor binding and epigenomic datasets.

The BiSA statistical overlap correlation value (OCV) represents a statistical summary value of the set of p -values calculated by the IntervalStat tool and reflects the overall correlation of two binding site datasets. IntervalStat calculates a p -value for each query region against the closest reference region within the given domain. It is designed to identify factors that target the same genomic locations. As described in examples in our previous study (Khushi et al., 2014) the OCV should be greater than 0.5 for partner factors, reflecting a statistically significant correlation between two binding patterns. For example the OCV for known partners, FOXA3 (query) to FOXA1 (reference) was 0.72 (Motallebipour et al., 2009). Similarly the OCV for CTCF (query) and SA1 (reference), which are known to co-locate on DNA, was 0.82 (Schmidt et al., 2010). Therefore the lower OCV for ER α -PR suggests that the majority of ER α and PR binding events are independent of each other, however, the OCV test does not challenge the biological

co-occurrence of binding of the two factors on the reported regions where IntervalStat reports a statistically significant p -value.

A consistent overlap was found both proximal and distal to gene promoters (Table 3). It is acknowledged that gene expression is regulated through interaction at a number of cis-regulatory elements, which includes promoters and enhancers. Moreover, enhancers can spread over a range of distances from the TSS. Therefore, the detection of binding sites over a range of distances and locations is to be expected (Bulger & Groudine, 2011; Calo & Wysocka, 2013). This spatial correlation between the two factors is identified as statistically significant by Monte Carlo simulation using BITS, Relevant Distance, Absolute Distance, Jaccard and Projection tests using Genometricorr. Therefore, the regions from the two factors are found in close proximity more often than expected by chance although they do not exactly overlap. Therefore the consistent OCV observed using various domains and statistically significant spatial convergence suggest that the consistent overlap may have biological significance. Although not all sites overlapped, many of the shared ER α and PR binding regions were highly statistically significant binding sites for both receptors, as determined by a strong p -value and low FDR value in MACS, suggesting that these are biologically valid binding regions for these receptors and that their overlap reflects converging function on a subset of gene targets.

In recent years a number of studies have published ER α binding regions in the MCF-7 cell line (Grober et al., 2011; Gu et al., 2010; Hu et al., 2010; Hurtado et al., 2008; Joseph et al., 2010; Schmidt et al., 2010; Tsai et al., 2010; Welboren et al., 2009). However only two studies have published ER α data in T47D cells (Gertz et al., 2012; Joseph et al., 2010). We chose to study the Gertz et al. (2012) dataset because using data from the Joseph et al. (2010) study we called only 1,817 peaks with FDR <5%, which can be an indication of low quality ChIP (Landt et al., 2012). On the other hand for the PR dataset, we did not employ the datasets published by Yin et al. (2012) because the experiment was performed with an antiprogestin (RU486) treatment, which would not be expected to elicit the same binding pattern as PR agonist, and lacked any control sample. MACS distributes read tags from the control sample along the genome to model Poisson distribution, and false discovery rate (FDR) is calculated by swapping control and ChIP samples. Therefore it is recommended for ChIP-seq studies to have an appropriate input control sample (Wilbanks & Facciotti, 2010). ENCODE guidelines also emphasise the importance of using a suitable control dataset to adjust for variable DNA fragment lengths (Landt et al., 2012).

There is a slight difference in the reported low-significance motifs for PR data between this report and the Clarke and Graham study (Clarke & Graham, 2012). The two most significant motifs (PRE are FOXA1) are the same in the two studies, however, Clarke and Graham found an NF1 half-site as one of the significant motifs and AP-1 sites as non-significant while in this study we found an AP-2 motif higher in significance than the NF1 motif (not shown). This minor difference is due to the difference in binding regions as Clarke and Graham published 6,312 PR bound regions in T47D cells by aligning to hg18 and using the ERANGE peak caller, however, in this study we reported 22,152 PR regions by aligning to hg19 assembly and using MACS as our peak caller.

The ER α -PR data was collected from two separate publications where the binding of each factor was studied by stimulation of T-47D cells with estrogen or progesterone independently. Therefore the focus of this study was to examine the correlation of ER α -PR binding patterns which revealed an interesting convergence on specific loci. We studied the association between common regions and nearby genes and found biologically relevant gene pathways. The Myc oncogene, which was most highly associated with binding sites common to ER α and PR, is a known target of both estrogen and progesterone and plays a key role in the normal breast and breast cancer (Curtis *et al.*, 2012; Hynes & Stoezel, 2009). PR itself is also regulated by both hormones and the PGR gene was highly associated with shared ER α and PR binding regions. Transcriptional regulation by estrogen and progesterone co-treatment in this cell model was not available, however it would be interesting to study the binding of the two factors under the influence of both stimuli (estrogen and progesterone) to observe the impact of converging ER α and PR regulation in comparison to individual stimulation.

CONCLUSION

In summary, we have evidence for a biologically relevant interplay between PR and ER α in a subset of binding sites in breast cancer cells. Our analysis demonstrated the utility of our previously published software BiSA (Khushi *et al.*, 2014), which has a comprehensive knowledge base, consisting of transcription factor binding sites and histone modifications collected from previously published studies. Using BiSA we identified that ER α and PR co-locate on a subset of binding sites. The BiSA statistical testing of overlap revealed a low overlap correlation value (OCV) suggesting that the two factors are not obligate cofactors. However, spatial correlation testing using Monte Carlo simulation by BITS, Relevant Distance, Absolute Distance, Jaccard and Projection tests by Genometricorr revealed a statistically significant correlation between the two factors. In addition, the discovery that ER α , FOXA1, PR, AP-2 and TEAD4 binding motifs are significantly enriched in regions that are bound by both ER α and PR suggests that their overlap is biologically relevant.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

MK was previously supported by Australian Postgraduate Award (APA) and Westmead Medical Research Foundation (WMRF) Top-Up scholarship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:

Australian Postgraduate Award (APA).

Westmead Medical Research Foundation (WMRF).

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Matloob Khushi conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Christine L. Clarke and J. Dinny Graham conceived and designed the experiments, reviewed drafts of the paper.

Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj.654#supplemental-information>.

REFERENCES

- Abdel-Hafiz HA, Horwitz KB. 2014.** Post-translational modifications of the progesterone receptors. *Journal of Steroid Biochemistry and Molecular Biology* **140**:80–89 DOI [10.1016/j.jsbmb.2013.12.008](https://doi.org/10.1016/j.jsbmb.2013.12.008).
- Augello MA, Hickey TE, Knudsen KE. 2011.** FOXA1: master of steroid receptor function in cancer. *EMBO Journal* **30**:3885–3894 DOI [10.1038/emboj.2011.340](https://doi.org/10.1038/emboj.2011.340).
- Ballare C, Castellano G, Gaveglia L, Althammer S, Gonzalez-Vallinas J, Eyraes E, Le Dily F, Zaurin R, Soronellas D, Vicent GP, Beato M. 2013.** Nucleosome-driven transcription factor binding and gene regulation. *Molecular Cell* **49**:67–79 DOI [10.1016/j.molcel.2012.10.019](https://doi.org/10.1016/j.molcel.2012.10.019).
- Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, Noushmehr H, Lange CPE, Van Dijk CM, Tollenaar RAEM, Van Den Berg D, Laird PW. 2012.** Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nature Genetics* **44**:40–46 DOI [10.1038/ng.969](https://doi.org/10.1038/ng.969).
- Bernardo GM, Keri RA. 2012.** FOXA1: a transcription factor with parallel functions in development and cancer. *Bioscience Reports* **32**:113–130 DOI [10.1042/BSR20110046](https://doi.org/10.1042/BSR20110046).
- Bulger M, Groudine M. 2011.** Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**:327–339 DOI [10.1016/j.cell.2011.01.024](https://doi.org/10.1016/j.cell.2011.01.024).
- Bulun SE. 2014.** Aromatase and estrogen receptor alpha deficiency. *Fertility and Sterility* **101**:323–329 DOI [10.1016/j.fertnstert.2013.12.022](https://doi.org/10.1016/j.fertnstert.2013.12.022).
- Cadoo KA, Fournier MN, Morris PG. 2013.** Biological subtypes of breast cancer: current concepts and implications for recurrence patterns. *The Quarterly Journal of Nuclear Medicine and Molecular Imaging* **57**:312–321.
- Calo E, Wysocka J. 2013.** Modification of enhancer chromatin: what, how, and why? *Molecular Cell* **49**:825–837 DOI [10.1016/j.molcel.2013.01.038](https://doi.org/10.1016/j.molcel.2013.01.038).
- Cerami EG, Bader GD, Gross BE, Sander C. 2006.** cPath: open source software for collecting, storing, and querying biological pathways. *BMC Bioinformatics* **7**:497 DOI [10.1186/1471-2105-7-497](https://doi.org/10.1186/1471-2105-7-497).
- Chalbos D, Vignon F, Keydar I, Rochefort H. 1982.** Estrogens stimulate cell proliferation and induce secretory proteins in a human breast cancer cell line (T47D). *Journal of Clinical Endocrinology and Metabolism* **55**:276–283 DOI [10.1210/jcem-55-2-276](https://doi.org/10.1210/jcem-55-2-276).

- Chikina MD, Troyanskaya OG. 2012.** An effective statistical evaluation of ChIPseq dataset similarity. *Bioinformatics* 28:607–613 DOI [10.1093/bioinformatics/bts009](https://doi.org/10.1093/bioinformatics/bts009).
- Clarke CL, Graham JD. 2012.** Non-overlapping progesterone receptor cistromes contribute to cell-specific transcriptional outcomes. *PLoS ONE* 7:e35859 DOI [10.1371/journal.pone.0035859](https://doi.org/10.1371/journal.pone.0035859).
- Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Graf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, Group M, Langerod A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowitz F, Murphy L, Ellis I, Purushotham A, Borresen-Dale AL, Brenton JD, Tavare S, Caldas C, Aparicio S. 2012.** The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486:346–352 DOI [10.1038/nature10983](https://doi.org/10.1038/nature10983).
- Cyr AR, Kulak MV, Park JM, Bogachek MV, Spanheimer PM, Woodfield GW, White-Baer LS, O'Malley YQ, Sugg SL, Olivier AK, Zhang W, Domann FE, Weigel RJ. 2014.** TFAP2C governs the luminal epithelial phenotype in mammary development and carcinogenesis. *Oncogene* In Press.
- D'Abreo N, Hindenburg AA. 2013.** Sex hormone receptors in breast cancer. *Vitamins and Hormones* 93:99–133 DOI [10.1016/B978-0-12-416673-8.00001-0](https://doi.org/10.1016/B978-0-12-416673-8.00001-0).
- Favorov A, Mularoni L, Cope LM, Medvedeva Y, Mironov AA, Makeev VJ, Wheelan SJ. 2012.** Exploring massive, genome scale datasets with the GenometriCorr package. *PLoS Computational Biology* 8:e1002529 DOI [10.1371/journal.pcbi.1002529](https://doi.org/10.1371/journal.pcbi.1002529).
- Gertz J, Reddy TE, Varley KE, Garabedian MJ, Myers RM. 2012.** Genistein and bisphenol a exposure cause estrogen receptor 1 to bind thousands of sites in a cell type-specific manner. *Genome Research* 22:2153–2162 DOI [10.1101/gr.135681.111](https://doi.org/10.1101/gr.135681.111).
- Goecks J, Nekrutenko A, Taylor J, Galaxy T. 2010.** Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* 11:R86 DOI [10.1186/gb-2010-11-8-r86](https://doi.org/10.1186/gb-2010-11-8-r86).
- Grober OM, Mutarelli M, Giurato G, Ravo M, Cicatiello L, De Filippo MR, Ferraro L, Nassa G, Papa MF, Paris O, Tarallo R, Luo S, Schroth GP, Benes V, Weisz A. 2011.** Global analysis of estrogen receptor beta binding to breast cancer cell genome reveals an extensive interplay with estrogen receptor alpha for target gene regulation. *BMC Genomics* 12:36 DOI [10.1186/1471-2164-12-36](https://doi.org/10.1186/1471-2164-12-36).
- Gu F, Hsu HK, Hsu PY, Wu J, Ma Y, Parvin J, Huang TH, Jin VX. 2010.** Inference of hierarchical regulatory network of estrogen-dependent breast cancer through ChIP-based data. *BMC Systems Biology* 4:170 DOI [10.1186/1752-0509-4-170](https://doi.org/10.1186/1752-0509-4-170).
- Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010.** Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell* 38:576–589 DOI [10.1016/j.molcel.2010.05.004](https://doi.org/10.1016/j.molcel.2010.05.004).
- Hu M, Yu J, Taylor JM, Chinnaiyan AM, Qin ZS. 2010.** On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Research* 38:2154–2167 DOI [10.1093/nar/gkp1180](https://doi.org/10.1093/nar/gkp1180).
- Hurtado A, Holmes KA, Geistlinger TR, Hutcheson IR, Nicholson RI, Brown M, Jiang J, Howat WJ, Ali S, Carroll JS. 2008.** Regulation of ERBB2 by oestrogen receptor-PAX2 determines response to tamoxifen. *Nature* 456:663–666 DOI [10.1038/nature07483](https://doi.org/10.1038/nature07483).
- Hynes NE, Stoelzle T. 2009.** Key signalling nodes in mammary gland development and cancer: Myc. *Breast Cancer Research* 11:210 DOI [10.1186/bcr2406](https://doi.org/10.1186/bcr2406).

- Ishikawa H, Ishi K, Serna VA, Kakazu R, Bulun SE, Kurita T. 2010. Progesterone is essential for maintenance and growth of uterine leiomyoma. *Endocrinology* 151:2433–2442 DOI 10.1210/en.2009-1225.
- Joseph R, Orlov YL, Huss M, Sun W, Kong SL, Ukil L, Pan YF, Li G, Lim M, Thomsen JS, Ruan Y, Clarke ND, Prabhakar S, Cheung E, Liu ET. 2010. Integrative model of genomic factors for determining binding site selection by estrogen receptor-alpha. *Molecular Systems Biology* 6:456 DOI 10.1038/msb.2010.109.
- Kalkman S, Barentsz MW, Van Diest PJ. 2014. The effects of under 6 hours of formalin fixation on hormone receptor and HER2 expression in invasive breast cancer: a systematic review. *American Journal of Clinical Pathology* 142:16–22 DOI 10.1309/AJCP96YDQSTYBXWU.
- Kerpedjiev P, Frellsen J, Lindgreen S, Krogh A. 2014. Adaptable probabilistic mapping of short reads using position specific scoring matrices. *BMC Bioinformatics* 15:100 DOI 10.1186/1471-2105-15-100.
- Kharchenko PV, Tolstorukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology* 26:1351–1359 DOI 10.1038/nbt.1508.
- Khushi M, Liddle C, Clarke CL, Graham JD. 2014. Binding sites analyser (BiSA): software for genomic binding sites archiving and overlap analysis. *PLoS ONE* 9:e87301 DOI 10.1371/journal.pone.0087301.
- Kim JJ, Kurita T, Bulun SE. 2013. Progesterone action in endometrial cancer, endometriosis, uterine fibroids, and breast cancer. *Endocrine Reviews* 34:130–162 DOI 10.1210/er.2012-1043.
- Kittler R, Zhou J, Hua S, Ma L, Liu Y, Pendleton E, Cheng C, Gerstein M, White KP. 2013. A comprehensive nuclear receptor network for breast cancer cells. *Cell Reports* 3:538–551 DOI 10.1016/j.celrep.2013.01.004.
- Ladunga I. 2010. An overview of the computational analyses and discovery of transcription factor binding sites. *Methods in Molecular Biology* 674:1–22 DOI 10.1007/978-1-60761-854-6_1.
- Lal G, Contreras PG, Kulak M, Woodfield G, Bair T, Domann FE, Weigel RJ. 2013. Human Melanoma cells over-express extracellular matrix 1 (ECM1) which is regulated by TFAP2C. *PLoS ONE* 8:e73953 DOI 10.1371/journal.pone.0073953.
- Lam EW, Brosens JJ, Gomes AR, Koo CY. 2013. Forkhead box proteins: tuning forks for transcriptional harmony. *Nature Reviews Cancer* 13:482–495 DOI 10.1038/nrc3539.
- Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shores N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research* 22:1813–1831 DOI 10.1101/gr.136184.111.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9:357–359 DOI 10.1038/nmeth.1923.
- Layer RM, Skadron K, Robins G, Hall IM, Quinlan AR. 2013. Binary Interval Search: a scalable algorithm for counting interval intersections. *Bioinformatics* 29:1–7 DOI 10.1093/bioinformatics/bts652.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079 DOI 10.1093/bioinformatics/btp352.

- Lim B, Park JL, Kim HJ, Park YK, Kim JH, Sohn HA, Noh SM, Song KS, Kim WH, Kim YS, Kim SY. 2014. Integrative genomics analysis reveals the multilevel dysregulation and oncogenic characteristics of TEAD4 in gastric cancer. *Carcinogenesis* 35:1020–1027 DOI 10.1093/carcin/bgt409.
- Lin CY, Vega VB, Thomsen JS, Zhang T, Kong SL, Xie M, Chiu KP, Lipovich L, Barnett DH, Stossi F, Yeo A, George J, Kuznetsov VA, Lee YK, Charn TH, Palanisamy N, Miller LD, Cheung E, Katzenellenbogen BS, Ruan Y, Bourque G, Wei CL, Liu ET. 2007. Whole-genome cartography of estrogen receptor alpha binding sites. *PLoS Genetics* 3:e87 DOI 10.1371/journal.pgen.0030087.
- Lunter G, Goodson M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research* 21:936–939 DOI 10.1101/gr.111120.110.
- McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology* 28:495–501 DOI 10.1038/nbt.1630.
- Mesrouze Y, Hau JC, Erdmann D, Zimmermann C, Fontana P, Schmelzle T, Chene P. 2014. The surprising features of the TEAD4-Vgll1 protein–protein interaction. *ChemBioChem* 15:537–542 DOI 10.1002/cbic.201300715.
- Motallebipour M, Ameer A, Reddy Bysani MS, Patra K, Wallerman O, Mangion J, Barker MA, McKernan KJ, Komorowski J, Wadelius C. 2009. Differential binding and co-binding pattern of FOXA1 and FOXA3 and their relation to H3K4me3 in HepG2 cells revealed by ChIP-seq. *Genome Biology* 10:R129 DOI 10.1186/gb-2009-10-11-r129.
- Nakshatri H, Badve S. 2009. FOXA1 in breast cancer. *Expert Reviews in Molecular Medicine* 11:e8 DOI 10.1017/S1462399409001008.
- Obiorah IE, Fan P, Sengupta S, Jordan VC. 2014. Selective estrogen-induced apoptosis in breast cancer. *Steroids* 90:60–70 DOI 10.1016/j.steroids.2014.06.003.
- Orr N, Lemnrau A, Cooke R, Fletcher O, Tomczyk K, Jones M, Johnson N, Lord CJ, Mitsopoulos C, Zvelebil M, McDade SS, Buck G, Blancher C, Consortium KC, Trainer AH, James PA, Bojesen SE, Bokmand S, Nevanlinna H, Mattson J, Friedman E, Laitman Y, Palli D, Masala G, Zanna I, Ottini L, Giannini G, Hollestelle A, Ouweland AM, Novakovic S, Krajc M, Gago-Dominguez M, Castela JE, Olsson H, Hedenfalk I, Easton DF, Pharoah PD, Dunning AM, Bishop DT, Neuhausen SL, Steele L, Houlston RS, Garcia-Closas M, Ashworth A, Swerdlow AJ. 2012. Genome-wide association study identifies a common variant in RAD51B associated with male breast cancer risk. *Nature Genetics* 44:1182–1184 DOI 10.1038/ng.2417.
- Penault-Llorca F, Viale G. 2012. Pathological and molecular diagnosis of triple-negative breast cancer: a clinical perspective. *Annals of Oncology* 23(Suppl 6):vi19–vi22 DOI 10.1093/annonc/mds190.
- Pepke S, Wold B, Mortazavi A. 2009. Computation for ChIP-seq and RNA-seq studies. *Nature Methods* 6:S22–S32 DOI 10.1038/nmeth.1371.
- Salehnia M, Zavareh S. 2013. The effects of progesterone on oocyte maturation and embryo development. *International Journal of Fertility & Sterility* 7:74–81.
- Schmidt D, Schwalie PC, Ross-Innes CS, Hurtado A, Brown GD, Carroll JS, Flicek P, Odom DT. 2010. A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Research* 20:578–588 DOI 10.1101/gr.100479.109.

- Shao R, Cao S, Wang X, Feng Y, Billig H. 2014. The elusive and controversial roles of estrogen and progesterone receptors in human endometriosis. *American Journal of Translational Research* 6:104–113.
- Ström A, Hartman J, Foster JS, Kietz S, Wimalasena J, Gustafsson J-Å. 2004. Estrogen receptor β inhibits 17β -estradiol-stimulated proliferation of the breast cancer cell line T47D. *Proceedings of the National Academy of Sciences of the United States of America* 101:1566–1571 DOI 10.1073/pnas.0308319100.
- Tsai MJ, O'Malley BW. 1994. Molecular mechanisms of action of steroid/thyroid receptor superfamily members. *Annual Review of Biochemistry* 63:451–486 DOI 10.1146/annurev.bi.63.070194.002315.
- Tsai WW, Wang Z, Yiu TT, Akdemir KC, Xia W, Winter S, Tsai CY, Shi X, Schwarzer D, Plunkett W, Aronow B, Gozani O, Fischle W, Hung MC, Patel DJ, Barton MC. 2010. TRIM24 links a non-canonical histone signature to breast cancer. *Nature* 468:927–932 DOI 10.1038/nature09542.
- Wang L, Di LJ. 2014. BRCA1 and estrogen/estrogen receptor in breast cancer: where they interact? *International Journal of Biological Sciences* 10:566–575 DOI 10.7150/ijbs.8579.
- Wang D, Garcia-Bassets I, Benner C, Li W, Su X, Zhou Y, Qiu J, Liu W, Kaikkonen MU, Ohgi KA, Glass CK, Rosenfeld MG, Fu XD. 2011b. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* 474:390–394 DOI 10.1038/nature10006.
- Wang C, Mayer JA, Mazumdar A, Fertuck K, Kim H, Brown M, Brown PH. 2011a. Estrogen induces c-myc gene expression via an upstream enhancer activated by the estrogen receptor and the AP-1 transcription factor. *Molecular Endocrinology* 25:1527–1538 DOI 10.1210/me.2011-1037.
- Welboren WJ, Van Driel MA, Janssen-Megens EM, Van Heeringen SJ, Sweep FC, Span PN, Stunnenberg HG. 2009. ChIP-Seq of ERalpha and RNA polymerase II defines genes differentially responding to ligands. *EMBO Journal* 28:1418–1428 DOI 10.1038/emboj.2009.88.
- Wilbanks EG, Facciotti MT. 2010. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE* 5:e11471 DOI 10.1371/journal.pone.0011471.
- Woodfield GW, Chen Y, Bair TB, Domann FE, Weigel RJ. 2010. Identification of primary gene targets of TFAP2C in hormone responsive breast carcinoma cells. *Genes Chromosomes Cancer* 49:948–962 DOI 10.1002/gcc.20807.
- Xia Y, Chang T, Wang Y, Liu Y, Li W, Li M, Fan HY. 2014. YAP promotes ovarian cancer cell tumorigenesis and is indicative of a poor prognosis for ovarian cancer patients. *PLoS ONE* 9:e91770 DOI 10.1371/journal.pone.0091770.
- Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, Whitaker JW, Tian S, Hawkins RD, Leung D, Yang H, Wang T, Lee AY, Swanson SA, Zhang J, Zhu Y, Kim A, Nery JR, Urich MA, Kuan S, Yen CA, Klugman S, Yu P, Suknuntha K, Propson NE, Chen H, Edsall LE, Wagner U, Li Y, Ye Z, Kulkarni A, Xuan Z, Chung WY, Chi NC, Antosiewicz-Bourget JE, Slukvin I, Stewart R, Zhang MQ, Wang W, Thomson JA, Ecker JR, Ren B. 2013. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* 153:1134–1148 DOI 10.1016/j.cell.2013.04.022.
- Yadav BS, Sharma SC, Chanana P, Jhamb S. 2014. Systemic treatment strategies for triple-negative breast cancer. *World Journal of Clinical Oncology* 5:125–133 DOI 10.5306/wjco.v5.i2.125.

- Yin P, Roqueiro D, Huang L, Owen JK, Xie A, Navarro A, Monsivais D, Coon JSt, Kim JJ, Dai Y, Bulun SE. 2012.** Genome-wide progesterone receptor binding: cell type-specific and shared mechanisms in T47D breast cancer cells and primary leiomyoma cells. *PLoS ONE* 7:e29021 DOI [10.1371/journal.pone.0029021](https://doi.org/10.1371/journal.pone.0029021).
- Zhang Y, Liu T, Meyer CA, Eeckhoutte J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. 2008.** Model-based analysis of ChIP-Seq (MACS). *Genome Biology* 9:R137 DOI [10.1186/gb-2008-9-9-r137](https://doi.org/10.1186/gb-2008-9-9-r137).