

SOURCE OF UNALIGNED ChIP-Seq READS

Wilberforce Zachary Ouma
I56/77774/2009

Center for Biotechnology and Bioinformatics
University of Nairobi

A thesis submitted to the Board of Postgraduate Studies,
University of Nairobi, in partial fulfillment for the award
of

Master of Science in Bioinformatics

©2012

DECLARATION

The work reported herein is original and has not been presented for a degree program in any academic institution or university.

Wilberforce Zachary Ouma

Signature  Date 10/01/2013

APPROVAL

This thesis has been submitted for examination with our approval as the University Supervisors:

Prof. James O. Ochanda

Signature  Date 11/02/2013

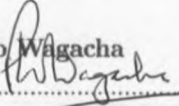
Centre for Biotechnology and Bioinformatics
University of Nairobi
P.O Box 30197- 00100
Nairobi, Kenya

Dr. Lynette Isabella Oyler

Signature  Date 01/2/2013

Center for Biotechnology and Bioinformatics
University of Nairobi
P.O Box 30197-00100
Nairobi, Kenya


Dr. Peter Waiganjo Wagacha

Signature  Date 7/2/13

School of Computing and Informatics
University of Nairobi
P.O Box 30197- 00100
Nairobi, Kenya

Dr. Erich Grotewold

Signature



Date

20-12-12

Department of Molecular Genetics

&

Center for Applied Plant Sciences

The Ohio State University

1060 Carmack Road

Columbus, Ohio 432302

United States of America

Acknowledgements

I do acknowledge God Almighty for the strength and grace He bestowed upon me while undertaking this research study. I am highly indebted to my research supervisors: Dr. Erich Grotewold for offering me an excellent opportunity to work in his world-class laboratory at the Ohio State University, United States of America. Erich, you have been a wonderful supervisor and an admirable mentor. I thank you, Dr. Lynette Oyier for your invaluable guidance in my research work, especially for offering constructive critique during the thesis writing process. I acknowledge constant support from Prof. James Ochanda through whom I got the opportunity of studying in one of the finest Bioinformatics academic and research programs in the country. Many thanks to Dr. Peter Waiganjo Wagacha for playing a vital role in fanning into flame my curiosity in computational sciences. I am sincerely grateful to my family for constant support, unconditional love and encouragement throughout my whole life and most importantly during my research work. I am immeasurably grateful to my Grandmother Jerusa Murwayi for instilling in me a great sense of discipline and hardwork; and for being an every-day source of inspiration. Finally, I thank the Federal Government of the United States of America for funding my research through the National Science Foundation grant.

Dedication

I dedicate this thesis to my lovely family: To Mary Wafula and Peninah Namussasi for seeing to it that I get the best education ever; to my siblings for making me smile even when the going gets tough; and to Jerusa Murwayi for being an epitome of hope and excellence.

Contents

- Contents** **iii**

- List of Figures** **vii**

- 1 Introduction** **1**
 - 1.1 Problem Statement 3
 - 1.2 Research Question 3
 - 1.3 Main Objective 4
 - 1.3.1 Specific Objectives 4
 - 1.4 Study Justification 4

- 2 Literature Review** **5**
 - 2.1 ChIP-Seq Protocol 5
 - 2.1.1 Chromatin enrichment and Sequencing 5
 - 2.1.2 Short read alignment 6
 - 2.1.3 Post-alignment Analysis 9
 - 2.2 Possible causes of read non-alignment 10
 - 2.2.1 Read sequence quality 10
 - 2.2.2 Systematic Bias and Nucleotide Composition 11
 - 2.2.3 Contamination 11

- 3 Methodology** **13**
 - 3.1 Data sets 13
 - 3.1.1 Raw Sequence Data sets 13
 - 3.1.2 Simulated Data 14

CONTENTS

3.2	Genome Alignment	15
3.3	Short Read Clustering	15
3.4	Nucleotide Database Search and Taxonomic Classification of reads	16
3.4.1	Metagenomics7	17
3.5	Statistical Analyses and Visualizations	18
4	Results	19
4.1	Short Reads Alignment	19
4.2	Short reads quality score	23
4.3	Short reads cluster sizes distribution	27
4.4	Taxonomic classification of unaligned reads	30
4.4.1	Higher rank taxonomic units	30
5	Discussion and Conclusion	38
5.1	Discussion	38
5.1.1	Variations in alignment proportions	38
5.1.2	Potentially legitimate reads	40
5.1.3	Contamination	41
5.2	Conclusion	42
	Appendix 1	44
	Bibliography	53

List of Abbreviations

BLAST	Basic Local Alignment Search Tool
BLAT	Basic Local Alignment Tool
BWT	Burrows-Wheeler Transform
ChIP-Seq	Chromatin Immunoprecipitation with Sequencing
CPU	Central Processing Unit
ENCODE	ENcyclopedia Of DNA Elements
FTP	File Transfer Protocol
GRNs	Gene Regulatory Networks
HPC	High Performance Computing
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
IgG	Immunoglobulin G
MAQ	Mapping & Assembly with Qualities
MEGAN	Metagenome Analyzer
modENCODE	Model organism ENcyclopedia Of DNA Elements
NCBI	National Center for Biotechnology Information NCBI

CONTENTS

NG7	Metagenomics 7
NGS	Next Generation Sequencing
OTUs	Operational Taxonomic Units
PCR	Polymerase Chain Reaction
RAM	Random Access Memory
SHRiMP	Short Read Mapping Package
SOAP	Short Oligonucleotide Alignment Program
SRA	Short Read Archive
TFBSs	Transcription Factor Binding Sites

List of Figures

1	Alignment proportions of ChIP-Seq reads in the five genomes . . .	20
2	Alignment proportions of reads from ChIP-Seq control experiments	22
3	correlation between the size of the ChIP-Seq reads data sets and proportion of unaligned reads	24
4	Sequence quality of aligned reads	25
5	Sequence quality of unaligned reads	26
6	Power-law distribution of frequency of read clusters in an Ara- bidopsis ChIP-Seq dataset	28
7	Power-law distribution of frequency of read clusters in a Maize ChIP-Seq dataset	29
8	Taxonomic classification of unaligned reads	32
9	Taxonomic classification of Human ChIP-Seq unaligned reads . .	33
10	Relative abundance of taxonomic units in ChIP-Seq data sets . .	36
11	Taxonomic classification of selected Human and Drosophila un- aligned reads	37

Abstract

Chromatin Immunoprecipitation with sequencing (ChIP-Seq) is an indispensable tool in understanding the dynamics and evolution of regulatory circuitry of prokaryotes and eukaryotes by mapping genome-wide transcription factor binding sites (TFBSs). Aligning short sequence reads to the reference genome is the first step in the ChIP-Seq data analysis pipeline. Significantly low alignment proportions would therefore have a negative impact on the identification of TFBSs and thereby undermine the process of deciphering true gene regulatory networks (GRNs).

Source of unaligned reads in ChIP-Seq studies has never been explored. This study employed a computational approach in determining source of unaligned reads from major model organisms: *Arabidopsis thaliana*, *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*; and *Zea mays*. The analysis of raw sequence reads obtained from the National Center for Biotechnology Information (NCBI) short read archive (SRA) revealed a significant level of contamination in ChIP-Seq unaligned reads with sequences of bacterial and metazoan origin, irrespective of the source of chromatin used for the ChIP-Seq studies. In agreement with other sequencing studies, results reported herein indicate that human sequences are the main source of contamination. Unexpectedly, however, was the observation that selected unaligned reads data sets contained significant numbers of legitimate reads that have mappable properties, but were missed out in the alignment process. This highlights a need to improve the currently utilized alignment algorithms.

Chapter 1

Introduction

Chromatin Immunoprecipitation (ChIP) followed by next generation sequencing (ChIP-Seq) is a revolutionary tool used in deciphering gene regulatory circuitry of an organism (Kim and Ren, 2006). The technology's main application is the identification of Transcription Factor Binding Sites (TFBSs) on a genome-wide scale, deciphering the role of histone modification in regulating gene expression and inference of genome-wide nucleosome positions (Johnson *et al.*, 2007). ChIP followed by microarray (ChIP-CHIP) is a similar technology employed in the identification of genome-wide regulatory sites through hybridization of immunoprecipitated DNA on a microarray chip rather than sequencing as is the case with ChIP-Seq (Johnson *et al.*, 2007). ChIP-Seq is however mostly preferred to ChIP-CHIP, due to its ability to identify TFBSs with much higher resolution. Similarly, ChIP-Seq has the advantage of having less noise coupled with potentially greater sensitivity compared to ChIP-CHIP (Kaufmann *et al.*, 2010a).

ChIP-Seq is a multi-step experimental and analytical procedure that begins primarily with fixation of *in vivo* cross-linked cellular DNA and proteins with formaldehyde, the most commonly used fixative. This is followed by chromatin isolation and shearing of the protein-DNA complexes, followed by sonication to generate DNA fragments cross-linked to proteins. Chemical and enzymatic methods of shearing chromatin are often employed. These complexes are then incubated with a specific antibody resulting in antibody-protein-DNA complexes, which are subsequently isolated, reverse cross-linked and purified to generate DNA fragments for sequencing on a massively-parallel platform (Johnson *et al.*,

2007; Kaufmann *et al.*, 2010a). SOLEXA/Illumina is currently the most commonly used sequencing platform for ChIP-Seq studies (Pepke *et al.*, 2009).

Next generation sequencing technologies generate tens of millions of short-read sequences per run used in subsequent analyses that are geared towards inference of gene regulatory networks. One paramount step involves mapping of the short reads to the reference genome to identify potential TFBSs. This is based on the premise that genomic regions that are putative TFBSs will exhibit significant short-read sequence enrichment compared to non-TFBSs (Pepke *et al.*, 2009). Technical challenges encountered in mapping reads to the genome are two-fold: Limited computational resources in terms of Central Processing Unit (CPU) time and space (memory); and the efficiency with which an alignment algorithm can map millions of short reads to the reference genome. The first drawback can be somewhat surmounted by the use of High Performance Computing infrastructure (HPC) that is becoming increasingly available to the scientific community, either through institutional super computing infrastructure or Cloud Computing. Some of the Cloud services can be provided at the commercial level (Amazon Cloud, (<http://aws.amazon.com/>)) and academic level for free (Data Intensive Academic Grid, <http://diagcomputing.org/>). On the other hand, increased efficiency in mapping of millions of short-read sequences is being addressed by the design and implementation of several algorithms that exhibit increased sensitivity and speed (Li *et al.*, 2008a; Trapnell and Salzberg, 2009). The most commonly used software for aligning short-read sequences to the reference genome are comprised of but not limited to bowtie (Langmead *et al.*, 2009), Short Oligonucleotide Alignment Program (SOAP) (Li *et al.*, 2008b), MAPPING with Quality scores program (MAQ) (Li *et al.*, 2008a) and SHort Read Mapping Package (SHRiMP) (Rumble *et al.*, 2009). It is however important to note that no single application stands out as the 'gold standard' in short-read alignments, since the process (of alignment) is confounded by many dynamics that range from sample preparation to the sequencing technology used. It is also typical in ChIP-Seq data analysis to use uniquely mapped reads for identification of enriched genomic regions. This can significantly reduce the fraction of reads used in the analysis, since almost half the number of reads can map to multiple genomic regions, at least in plant genomes like *Arabidopsis thaliana*, *Oryza sativa* and *Zea mays* (Kaufmann *et al.*,

2010a). The implication of this is ChIP-Seq experiments need to generate a significant amount of reads that can be uniquely-mapped to the reference genome, but practically this is rarely the case (Guertin and Lis, 2010; Kaufmann *et al.*, 2010a).

While ChIP-Seq studies make use of uniquely-mapped reads to the reference genome in data analysis, the scientific community is yet to explore the biological significance of unaligned reads. Theoretically, these reads should align to the reference genome from which they have been sequenced. It is of interest to provide not just a technical explanation of read non-alignment but also to uncover biological relevance, if any, in the unaligned reads. This study focused on determining the source of unaligned short read sequences generated on SOLEXA/Illumina sequencing platform in ChIP-Seq experiments aimed at identifying Transcription Factor Binding Sites in *Arabidopsis thaliana*, *Zea mays*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens*. The afore-mentioned genomes have been selected due to the availability of raw Illumina-generated ChIP-Seq reads in the NCBI Short Read Archive (SRA) database. In addition, their reference genomes are known to be almost fully sequenced and well annotated.

1.1 Problem Statement

A significantly low proportion of ChIP-Seq reads align to their respective reference genomes (Guertin and Lis, 2010; Kaufmann *et al.*, 2010a). The implication of this is that not all of the sequenced reads are used to answer the biological question, and as a result the full potential of sequenced reads is not exploited.

1.2 Research Question

The problem of low read alignment in ChIP-Seq studies has not been previously addressed. This project therefore aims to identify the source of unaligned reads in ChIP-Seq experiments.

1.3 Main Objective

To explore and determine sources of unaligned ChIP-Seq short read sequences in *Arabidopsis thaliana*, *Zea mays*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens* ChIP-Seq experiments.

1.3.1 Specific Objectives

1. To determine ChIP-Seq short sequence reads alignment proportions in different experiments in the afore-mentioned genomes.
2. To determine the Operational Taxonomic Units (OTUs) of the unaligned reads and compare the trend across different experiments, laboratories and genomes.

1.4 Study Justification

Efficient mapping of short-read sequences to their respective reference genomes is a paramount step in ChIP-Seq data analysis. Grotewold laboratory and other laboratories (Kaufmann *et al.*, 2010b) carrying out ChIP-Seq experiments have observed a characteristically low fraction of read alignment/mappability to the reference genomes. Despite this observation, no systematic study has been carried out to determine the source of the unaligned reads in these experiments. Determining the cause and source of unaligned reads will guide researchers on better ChIP-Seq experimental design and improved quality control measures during the data preprocessing steps. A thorough understanding of the causes of the observed low fractions of aligned reads will contribute to the development of experimental techniques and analytical procedures that maximize the fraction and efficiency of mapping reads to the reference genome. This will result in the identification of TFBSs that would otherwise be missed out in the analysis. It is important for the scientific community involved in next-generation sequencing studies, especially in ChIP-Seq studies, to know why reads are not mapping to the reference genome in order to design better quality sequencing experiments (Kaufmann *et al.*, 2010a).

Chapter 2

Literature Review

2.1 ChIP-Seq Protocol

Chromatin Immunoprecipitation (ChIP) is a technique that is used to localise genome-wide binding sites for DNA-binding proteins of interest (Kaufmann *et al.*, 2010a). Such DNA-associated proteins include transcription factors (TF), DNA Polymerases and chromatin-associated histones. The technique exploits the ability of cross-linking these proteins to the DNA on which they bind *in vivo*. In the experimental set up, DNA-protein complexes are purified by immunoprecipitation using antibodies specific to the bound protein of interest. In order to generate the genome-wide protein binding profiles, whole genome tiling-arrays (ChIP-CHIP) or next generation sequencing (NGS) technologies (ChIP-Seq) are employed (Barski and Zhao, 2009).

2.1.1 Chromatin enrichment and Sequencing

The basic ChIP-Seq procedure begins with crosslinking *in vivo* DNA-protein interactions in the tissue of interest using a fixative like formaldehyde. This is followed by isolation of the nuclei and shearing of chromatin, preferably by sonication. The DNA-protein complexes are then immunoprecipitated by incubation with a specific antibody followed by the isolation of DNA by reverse cross-linking using proteinase K digestion and purification. Generally, it is good practice to

check the quality of the treatment sample using quantitative PCR before proceeding to the sequencing step. This process tests the enrichment of positive control sequences, which are simply genomic regions known to bind the TF of interest. The qPCR step is then followed by the generation of sequencing libraries of small (200 - 500 bps) fragments flanked by adaptors, which are then subjected to high-throughput sequencing by any of the available platforms like Illumina (Johnson *et al.*, 2007; Kaufmann *et al.*, 2010a). Although not the only platform, Illumina is currently the most commonly used sequencing platform for ChIP-Seq studies, resulting in millions of short sequence reads (hereafter referred to as reads) (Pepke *et al.*, 2009).

In order to identify the genomic regions recognized by a specific TF, reads are mapped to a reference genome to generate alignment profiles represented as peaks from which binding sites, DNA-binding motifs and TF target genes are determined. This is the first step in the ChIP-Seq data analysis pipeline. Correct peaks generated from alignment profiles are selected after comparison with peaks generated from control ChIP-Seq experiments, which can be one of the three alternatives: 1) Genomic DNA not obtained by immunoprecipitation; 2) DNA obtained by immunoprecipitation of chromatin from an individual lacking the particular TF (e.g., because of a mutation); and 3) DNA obtained by a mock immunoprecipitation of genomic DNA by an idiotypic control, such as commercial Immunoglobulin G (IgG). Read alignment is therefore one of the most important steps in the analysis pipeline.

2.1.2 Short read alignment

Mapping millions of short reads to the reference genome is one of the major challenges in NGS technologies. Technical challenges are associated with computational time and memory requirements for mapping short reads, while the analytical challenge is associated with the efficiency with which reads are mapped to the reference genome, otherwise referred to as read mappability (Trapnell and Salzberg, 2009). In order to circumvent these challenges, several alignment algo-

gorithms with different working techniques have been developed and optimized for aligning short-read sequences to the reference genome (Li *et al.*, 2008a,a; Rumble *et al.*, 2009). Unlike conventional alignment algorithms like Basic Local Alignment Search Tool (BLAST)(Altschul *et al.*, 1990) and Basic Local Alignment Tool (BLAT) (Kent, 2002), these newly developed algorithms that make use of indexing and heuristic searches have been designed to deal with the heavy computational load of aligning millions of short reads in significantly less CPU time compared to BLAST or BLAT, which can take hundreds or thousands of CPU hours on large datasets. (Trapnell and Salzberg, 2009).

Major alignment software can be classified into three major categories (Li *et al.*, 2008a). The first software category is based on hash tables as a way of indexing. In this approach, the first sub-category of software uses reads for creating a hash table and a reference genome for scanning the table, while another sub-category of software uses the reference genome for hashing and reads for scanning. MAQ, ZOOM and SeqMap make use of the former approach while the latest version of SHRiMP (Rumble *et al.*, 2009) employs the latter approach. A hash table is simply a data structure that is used to effectively index non-ordered data in order to achieve rapid searching.

A second category of software makes use of an algorithm developed by Burrows and Wheeler called Block-sorting algorithm, commonly referred to as the Burrows-Wheeler Transform (BWT) algorithm (Li *et al.*, 2009). Originally developed for data compression purposes, BWT has been implemented in several programs some of which include BWA (Li *et al.*, 2008a), bowtie (Langmead *et al.*, 2009), and SOAP2 (Li *et al.*, 2009). Both bowtie and BWA have been modified to include the Smith-Waterman algorithm, a dynamic programming algorithm that keeps track of all possible sequence alignments and outputs optimal alignments based on a specified scoring function. In addition, a more robust indexing strategy known as the FM-index has been incorporated into bowtie making it a computationally fast and memory efficient aligner.

The BWT algorithm transforms a string of characters by performing reversible permutations on the order of characters without changing their individual values. In the simplest form, the algorithm takes a string of characters with an end-of-file

(EOF) pointer/character as the last character and repositions the pointer in all possible positions in the string. The result is a table of all possible rotations of the string, which are then sorted in a lexicographic order and the last column taken as the transformed string. This in turn result to an easily encoded output that is easily reversible to the original input string. This algorithm enables large texts to be searched efficiently in a small memory footprint, which happens to be one of the strengths of bowtie. Note that BWT algorithm in bowtie is used mainly in the read indexing step in order to facilitate alignment. Bowtie additionally carries out a greedy, randomized, depth-first search through the space of all possible alignments. Apart from being the most commonly used program in the community, bowtie boasts of relatively high speed and memory efficiency in the alignment process compared to other whole-genome alignment programs, a vital feature for aligning reads from hundreds of raw ChIP-Seq data sets (Li and Homer, 2010; Schbath *et al.*, 2012; Trapnell and Salzberg, 2009). The last category of alignment algorithms uses merge-sort algorithm that sorts both the reference genome and the reads. Slider (Malhis *et al.*, 2009) is an implementation of this algorithm, specifically for the alignment of reads from the Illumina platform.

Several factors determine the choice of an alignment algorithm and these could range from CPU time and memory requirements to the efficiency of the alignments. Efficiency in this context refers to the fraction of reads that can be mapped to the reference genome by a specific algorithm. As mentioned above, bowtie has been reported to be a fast and a memory 'friendly' alignment program compared to other programs (Trapnell and Salzberg, 2009), while SHRiMP has been reported to take the longest CPU time in alignments albeit with increased read mappability. Trapnell observed at while there is increased time and memory efficiency in the newly-developed alignment programs, they achieve this at the cost of alignment specificity (Trapnell and Salzberg, 2009).

2.1.3 Post-alignment Analysis

Uniquely-mapped short reads are used in the post-alignment analyses, which include identification of tag/read enriched regions represented as alignment peaks. 'True' peaks are identified as those that have heights above a specified threshold, in most cases those whose heights are larger than the heights of reads generated from either an input control genomic DNA (non-immunoprecipitated genomic DNA), IgG immunoprecipitated DNA, or DNA from a sample of a mutant organism that does not express the TF of interest (Landt *et al.*, 2012). Alternatively, a reference background can be simulated and modelled, given a reference genome sequence, and used as an analytical control (Kaufmann *et al.*, 2010a). Genomic regions with statistically significant higher peaks compared to control sample peaks are considered to be putative Transcription Factor Binding Sites (TFBSs). Other analyses include identification of TFBSs sequences (Cis-regulatory elements) and TF target genes. As mentioned before, unaligned reads are usually not included in the analysis pipeline.

Unlike the conventional ChIP-Seq analysis pipeline, work reported herein takes a trajectory that involves taxonomic classification of unaligned reads using Metagenomics7 (MG7) (Pareja-Tobes *et al.*, 2012). The advantage of MG7 system is its ability to utilize a graph database model for the storage of query reads and their respective BLAST hit results. The system specifically uses Neo4j technology, integrating the results of the analysis with the Bio4j technology (<http://www.bio4j.com>), thus allowing its access within the framework of the NCBI taxonomy tree. Neo4j (<http://neo4j.org/>) is an open-source graph database implementation in java that stores data structured in graphs rather than in tables. Bio4j is simply an implementation of Neo4j specifically for biological data. The graph structure organizes data of taxonomically-assigned reads in a way semantically equivalent to what it represents- nodes representing query reads and their respective taxonomic units each connected by an edge, allowing both complex phylogenetic querying and a fine data access granularity in such a way that access to the specific results of each sequence can be attained. Results of the MG7 taxonomic assignment of reads were exported in different data formats,

XML, CSV and Gexf (Graph exchange XML format), thereby enabling different approaches for visualizations.

2.2 Possible causes of read non-alignment

2.2.1 Read sequence quality

ChIP-Seq reads that align to multiple regions of the genome are usually excluded from the analysis pipeline since they are considered ambiguous and can therefore contribute to false positives (Kaufmann *et al.*, 2010a,b). Several factors seem to influence read mappability to the reference genome, ranging from library preparation to error profile of the sequencer. Kaufman and colleagues (Kaufmann *et al.*, 2010a) demonstrated that there is a clear variation in read mappability to the *Arabidopsis thaliana* reference genome when libraries with different qualities are used. Library quality in this context implies the number of reads generated that have the adapter sequences used in library preparation, rendering such reads unmappable to the reference genome. Other possible factors include sequencing technology biases, for instance Illumina's bias of under-representation of read mappability in less GC rich regions, Cytosine-Phosphate-Guanine (CpG) islands, promoter regions, and poly[A] regions (Cheung *et al.*, 2011), (Nakamura *et al.*, 2011), (Dohm *et al.*, 2008), (Kaufmann *et al.*, 2010a). The base calling quality of reads also plays a role in read mappability. It has been observed that the base-call quality reduces in the 5' to 3' direction of the read, and this could be due to the reduced processivity and fidelity of the DNA polymerase in the sequencing process (Nakamura *et al.*, 2011). In order to address this issue, quality checks on reads can be carried out to remove reads that do not fit a set of determined quality standards before alignment to the reference genome. Similarly, one could trim sections of the reads that have poor base call quality (in most cases the 3' region) before aligning to the genome, however care must be taken to avoid a significant reduction in the read length to a level where the read will be too short to map uniquely to the genome. FASTQC (<http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>) is a program that has been developed for Illumina short read quality analysis and can be used to identify sequences that need to be trimmed before alignment.

Similarly, SOAP2 is able to iteratively trim reads with low quality bases in each alignment process until the read maps to the genome (Kaufmann *et al.*, 2010a).

2.2.2 Systematic Bias and Nucleotide Composition

Cheung and colleagues have recently reported that reads generated from ChIP-Seq input DNA do not exhibit a uniform distribution of mappability across the *C. elegans* and Human genomes, but rather the distribution of mappability is correlated with GC content in 1kb windows of the reference genome (Cheung *et al.*, 2011). There seemed to be a positive correlation between read mappability and increasing GC content of up to around 45% GC content, then the relationship became inverse, as reported in the same study (Cheung *et al.*, 2011). This was observed in reads generated from both input DNA and TF-immunoprecipitated (treatment) samples. Other studies have reported the correlation between GC content and read mappability to the reference genome (Dohm *et al.*, 2008; Nakamura *et al.*, 2011). A more recent study implicates CpG islands, promoter regions and 5' UTR regions as regions where there is reduced read mappability in comparison to other genomic regions. The exact mechanism by which nucleotide composition affects read mappability is yet to be determined and it is not known whether nucleotide bias is introduced at the library preparation, PCR, sequencing or alignment step.

2.2.3 Contamination

Sample DNA can be contaminated when foreign DNA is introduced before the adapter ligation step while preparing sequencing libraries (Kaufmann *et al.*, 2010a; Landt *et al.*, 2012). Similarly, the very step of adapter ligation can introduce adapter concatemers some of which will remain in the sample DNA even after size selection by gel filtration (Kaufmann *et al.*, 2010a). Determining the proportion of reads representing the adapter sequence can be used as a way of identification and removal of such sequences before the alignment of reads to the reference genome. A more efficient way of dealing with adapter concatemers

would be to optimize the ratio of sample DNA molecules to adapter molecules during the sample ligation step (Kaufmann *et al.*, 2010a).

One other possible source of contamination in sequencing experiments could be human DNA. This is could presumably be due to either contamination from researchers performing the experiment and/or extensive contamination of public sequence databases with human DNA, as shown by an in depth sequence search in NCBI, Ensembl, JGI, and UCSC carried out by Longo and colleagues (Longo *et al.*, 2011).

A review of literature established that no systematic study to determine the source of unaligned reads has been carried out.

Chapter 3

Methodology

3.1 Data sets

3.1.1 Raw Sequence Data sets

Short sequence reads from SOLEXA/Illumina sequencing platform were obtained from the NCBI (SRA) database using Linux `wget` - a non-interactive GNU command-line tool for file retrieval using HTTP, HTTPS and FTP Internet protocols. In order to enable comparison of results from analysis of different data sets and experiments, reads from experiments that were aimed at understanding transcriptional regulation due to transcription factor binding (and occasionally RNA Polymerase II binding) were downloaded. ChIP-Seq data sets of short sequence reads from experiments in *A. thaliana*, *C. elegans*, *D. melanogaster* and *Homo sapiens* were downloaded in an archived file format (NCBI .sra format) and subsequently converted to the FASTQ file format using the NCBI SRA toolkit 'fastq-dump' utility. In addition, data generated from Grotewold Laboratory for *Z. mays* transcription factors binding were included in the study for analysis.

All raw sequence reads in Human ChIP-Seq experiments were obtained from the Human Encyclopedia of DNA Elements (ENCODE) project aimed at deciphering functional elements in the human genome (Birney *et al.*, 2007). In this respect, these raw data sets contained reads from individual ENCODE ChIP-Seq experiments involved in determining DNA-binding profiles of the following

Transcription Factors: (a) Erythroblast Transformation Specific (ETS) family of transcription factors; (b) T-cell acute lymphoblastic leukemia1 (TAL1) binding in hematopoietic cell lineages; (c) Signal Transducer and Activator of Transcription1(STAT1) binding in Human HeLa S3 cells; and (d) RNA Polymerase II (PolII) binding in HeLa S3. Additionally, sequence data from the model organism ENCyclopedia Of DNA Elements (modENCODE) project (<http://www.modencode.org/>) involved in functional annotation of DNA elements in the worm model organism *C. elegans* were included in this study. These data sets contained ChIP-Seq reads from studies aimed at reconstructing gene regulatory circuitry involving LIN, Helix-loop-helix (HLH), Erythroid-like Transcription factor family (ELT), and Egg-laying (EGL) TFs, as well as POLII binding patterns. *C. elegans* raw data sets were however not restricted to those obtained from the modENCODE project. The *D. melanogaster* data sets on the other hand contained raw sequence reads from a study examining the Drosophila Heat Shock Factor (HSF) binding (Guertin and Lis, 2010). Raw data sets of the model plant organism *A. thaliana* were obtained from the following gene regulatory studies: The MADS-domain transcription factors APETALA1 (AP1) (Kaufmann *et al.*, 2010b), APETALA2 (AP2) and SEPALLATA3 (Kaufmann *et al.*, 2009)- key regulators of Arabidopsis floral organs development; and LEAFY (LFY) TF. The other plant-related ChIP-Seq data sets (which were generated in-house) were derived from Pericarp Color1 (P1) and Knotted1 (KN1) TF binding profiles in maize (Morohashi *et al.*, 2012). The description of each of the data sets/ChIP-Seq runs analyzed in this study is found in Appendix 1

3.1.2 Simulated Data

Genome-simulated data sets of 50 nucleotide sequence reads were included in the analysis as a positive controls. These synthetic data sets comprised of reads simulated from *C. elegans* (referred to as *C. elegans*-simulated) and *A. thaliana* (referred to as *tair9*-simulated) genomes- representing metazoan and plant genomes, respectively. Simulation of reads was performed using MetaSim, a sequence simulation tool for genomics and metagenomics analysis developed by Ritcher and

colleagues (Richter *et al.*, 2008). Each synthetic data set contained approximately five million short reads. A subset of aligned ChIP-Seq reads was also included in the analysis as a positive control.

3.2 Genome Alignment

Rather than pooling data from different runs in the same experiment, we analysed data sets independently in order to identify any variation that might exist within experiments. Sequence quality was first assessed using FastQC and reads were subsequently mapped to the reference genome using bowtie (Langmead *et al.*, 2009) in a multi-thread mode running on a multicore cluster. Mapping reads to their respective genomes served two purposes: (1) to observe variations in the amount of both aligned and unaligned reads across different ChIP-Seq runs, experiments and organisms; and (2) to obtain unaligned reads for further analysis in determining their provenance. Alignment was performed in parallel mode on an HP Intel Xeon Cluster with a total of 8,328 cores, 12 cores per node and 48 gigabytes (GB) of random access memory (RAM) per node or 4.0 GB per core. Unlike the conventional ChIP-Seq downstream data analysis process, reads that did not align to their respective reference genome were identified for a detailed analysis as reported in the following sections.

3.3 Short Read Clustering

Since the number of unaligned reads poised for downstream analyses was in hundreds of millions, it was imperative to reduce read redundancy inherent in the sequencing process by performing a read clustering procedure using UCLUST (Edgar, 2010) in order to obtain representative reads from each cluster for analysis. Additionally, the clustering process was carried out as a way of determining the distribution of the size of the clusters with an aim of comparing this distribution with the distribution of clusters generated from positive control data sets.

Clustering was an iterative process involving grouping sequences that meet a predefined sequence similarity score of 75%. Thus, sequences that exhibited a similarity of 75% or more were grouped in one cluster. The optimal sequence similarity score was determined after carrying out a series of pilot hierarchical clustering procedures, on a subset of the data set, at an increasing similarity score gradient. Simply, a subset of maize ChIP-Seq data set was clustered using UCLUST at increasing similarity scores of 70%, 75%, 80%, 85% and 90%; and then the distribution of the number of clusters in each similarity score determined. The optimal similarity score was determined by assessing the number of clusters generated in each score category and the average size of short read sequences. In this assessment, a similarity score of 75% was chosen to cluster all data sets. Although the number of clusters generated by this score was close to that generated by the 70% score, the 75% score prevents formation of spurious clusters compared to the 70% score, especially due to the nature of the short reads (50nt).

Representative sequences from each cluster, hereafter referred to as seeds, were obtained for taxonomic assignment. In addition, in order to develop an appropriate background model to determine the distribution of unaligned reads into clusters, the simulated 50 nt long reads from *A. thaliana* and *C. elegans* reference genomes were included in the analysis, as well as aligned reads from selected data sets. These data sets served as positive controls in subsequent analyses.

3.4 Nucleotide Database Search and Taxonomic Classification of reads

NCBI's pre-formatted nucleotide (nt) database was downloaded to the local compute cluster for ease of BLAST search. The nt database search was performed using NCBI's BLAST+ stand-alone tool (<http://www.ncbi.nlm.nih.gov/books/NBK1763/>) in a multi-thread mode and the BLAST results generated in BLASTXML format. This was followed by assignment of reads into their respective Operational Taxonomic Units (OTU) based on their BLAST hits/results using MEGAN4 (Huson *et al.*, 2011)- a metagenome analyzer; and Metagenomics7 (Pareja-Tobes

et al., 2012)(MG7) - an open source system for massive analysis of sequences from metagenomics samples. Both systems carried out taxonomic classification of short reads using NCBI's taxonomic tree and BLASTXML results. MG7 was however preferred due to its ability to carry out assignment of reads to their respective OTU on a high-throughput scale.

3.4.1 Metagenomics7

Metagenomics7 (Pareja-Tobes *et al.*, 2012) (hereafter simply referred to as MG7) was used on a cloud computing platform (<http://aws.amazon.com/ec2/>) to solve the problem of massive data analysis thereby dealing with the issues of analysing large data sets in a more efficient manner. This was particularly important in this study since approximately 1 terabyte of sequence data was analyzed. The assignment of taxonomic origin for each short sequence reads using MG7 was based on massive BLAST similarity analysis. This application includes a massive nucleotide BLAST program (BLASTN) that effects nucleotide searches against all nt databases. However, in this study this step was obviated since the BLAST results were obtained by a command-line nt database search tool. OTU assignment using MG7 therefore involved just the use of BLAST results after performing a nt database search.

Metagenomics7 implemented two different paradigms for the taxonomic assignment of reads: (i) Best Blast Hit (BBH) paradigm; and (ii) Lowest Common Ancestor (LCA) paradigm. Unlike the BBH paradigm in which a query read is assigned to the taxonomic unit of its 'best' hits, the LCA paradigm assigns the read to the taxonomy node corresponding to the Lowest Common Ancestor of the BLAST hits that passed the filter previously specified. MG7 provides the possibility of choosing different parameters to set the threshold for filtering the BLAST hits. In this study we chose (query-hit) alignment identity percentage and the query coverage percentage to filter the BLAST hits. Only BLAST hits with more than 95% identity within the BLAST High-scoring segment pairs (HSP) and with more than 95% of query coverage were selected and included in the taxonomic classification of query reads.

A parallel BLAST search and taxonomic classification of reads was carried out on the two data sets of genome-simulated reads and a subset of aligned reads to serve as a positive control. It was expected that almost all the reads in these control data sets would be assigned to the same taxonomic units as the read source

3.5 Statistical Analyses and Visualizations

Stacked bar plot representations of both alignment proportions and higher rank taxonomic units; as well as power-law distribution of the sequence clusters were generated on R statistical environment (<http://www.r-project.org/>) using a grammar for graphics (ggplot) package (<http://had.co.nz/ggplot2>). Network visualization of lower rank taxonomic units in each of the ChIP-Seq data sets was generated on gephi (<http://gephi.org/>), an interactive graph visualization and exploration platform for networks and complex systems.

Chapter 4

Results

4.1 Short Reads Alignment

Illumina-generated short sequence reads from ChIP-Seq studies aimed at identifying various TFBSs in the afore-mentioned five organisms were subjected to genome alignment using bowtie. Figure 1 is a bar plot showing percentages of uniquely-aligned, multiple-aligned and unaligned reads in each of the ChIP-Seq runs of the five organisms analyzed. We observed large variations in proportions of aligned reads in all the five genomes, with some genomes like *C. elegans*, *A. thaliana* and maize having runs with as low as 10% of reads uniquely aligning to their reference genomes, while some runs exhibited as high as 80% of reads aligning uniquely.

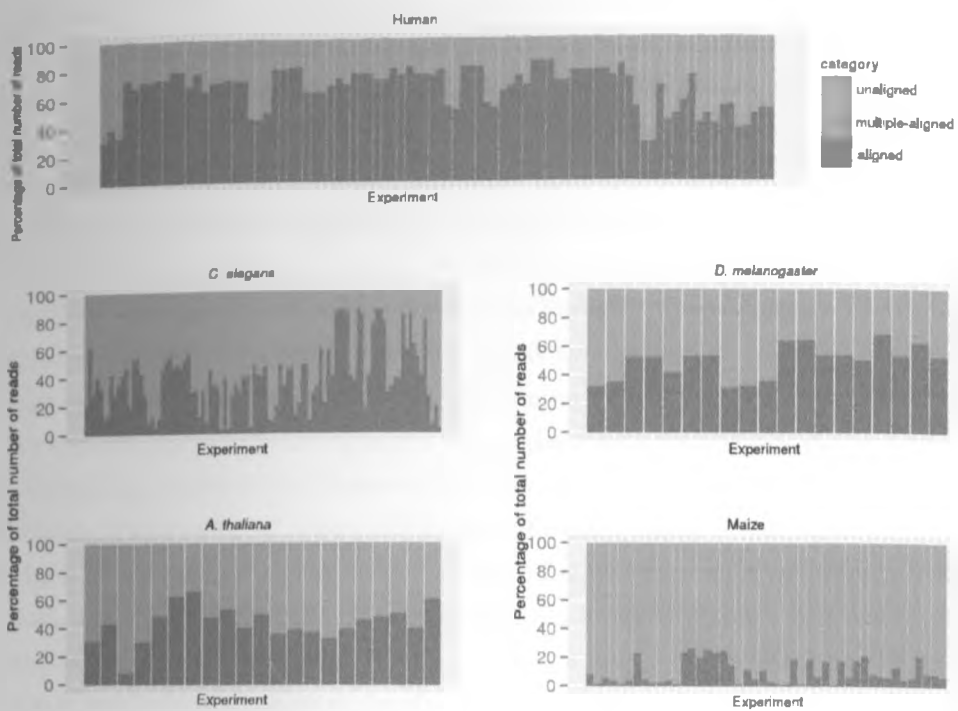


Figure 1: Alignment proportions of ChIP-Seq reads in the five genomes: each bar represents a ChIP-Seq run.

In human ChIP-Seq data sets for instance, which were predominantly from the ENCODE project (Gerstein *et al.*, 2012), alignment proportions varied considerably across different runs, with some runs having as low as 30% of uniquely aligned reads while others with as high as 80% of uniquely aligned reads (figure 1). These data sets also exhibited multiplicity in alignments in which, on average, 20% of the reads aligned to multiple regions of the genome. This trend was also observed in data sets from *C. elegans*, *A. thaliana*, and maize ChIP-Seq runs. The proportions of aligned reads in these organisms varied between different ChIP-Seq experiments within the same laboratory, albeit with relatively lower proportions of uniquely aligned reads compared to the human data sets (figure 1).

We then compared these alignment proportions with those generated in ChIP-Seq control experiments that involved the use of either non-immunoprecipitated chromatin, referred to as input DNA; or chromatin mock-immunoprecipitated with IgG. When analyzing input DNA reads, a higher proportion of aligned reads and less variation between samples was observed for both human and maize experiments, as depicted in figure 2, although maize had a significantly higher proportion of multiple aligned reads (multi-reads), possibly a consequence of the high proportion of paralogous genes (Schnable *et al.*, 2009) and the high abundance of transposons in the maize genome (Kronmiller and Wise, 2009; SanMiguel *et al.*, 1998; Schnable *et al.*, 2009). Unlike human and maize ChIP-Seq experiments, runs analysed from *C. elegans*, *D. melanogaster* and *A. thaliana* experiments did not involve use of input DNA as a negative control, instead a uniform distribution of background noise was simulated and used as a ChIP-Seq negative control.

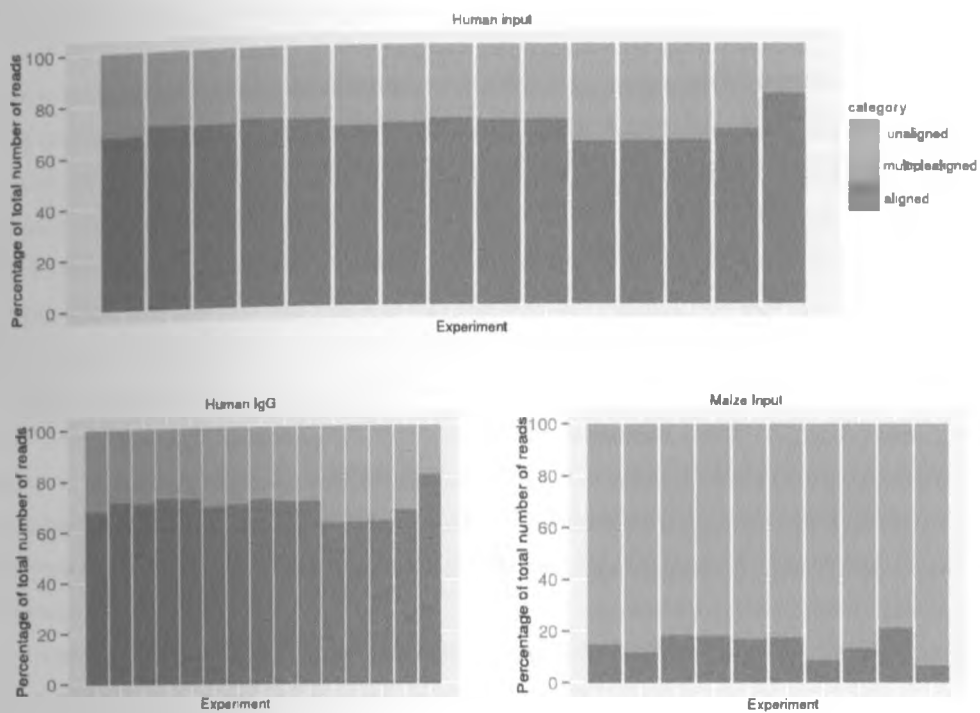


Figure 2: Alignment proportions of reads from ChIP-Seq control experiments: each bar on the x-axis represents a ChIP-Seq run. Multiplicity of alignment is significantly observed in the maize input ChIP-Seq data sets

4.2 Short reads quality score

We investigated whether the amount of unaligned reads correlates with the number of sequenced reads in each ChIP-Seq run. We observed no correlation between the size of ChIP-Seq data sets and the amount of unaligned reads in the five genomes analysed (figure 3), implying that more sequenced reads in a ChIP-Seq experiment do not necessarily guarantee better alignment proportions.

One possible premise for the high proportion of unaligned reads is poor sequence quality generated by the sequencing process. In order to determine the role played by sequence quality in the alignment process, we analysed and compared the quality of bases in both aligned and unaligned raw sequence reads using FastQC, a quality control tool for high throughput sequence data. The average per base and sequence quality scores for both aligned (figure 4) and unaligned (figure 5) reads were strikingly similar, albeit with reduced base quality at the 3 end of the sequences.

Moreover, we observed that on average, potential adapter sequences accounted for less than 0.5% of the unaligned reads, which is a negligible proportion to account for the high unaligned proportions observed.

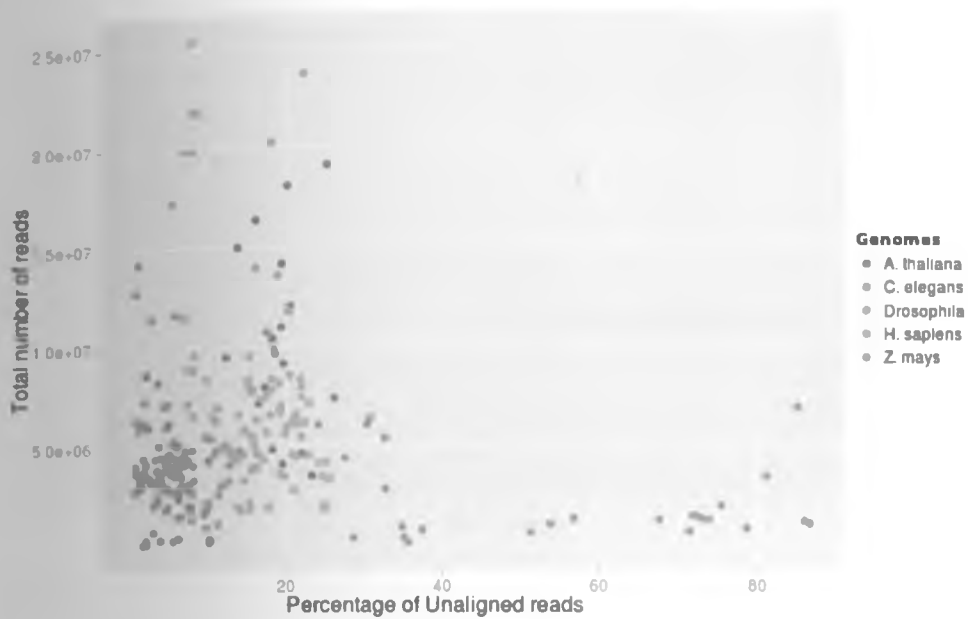


Figure 3: Scatter plot showing no correlation between the size of the ChIP-Seq reads data sets (number of sequences in a ChIP-Seq run) and the proportion of unaligned reads in the ChIP-Seq run.

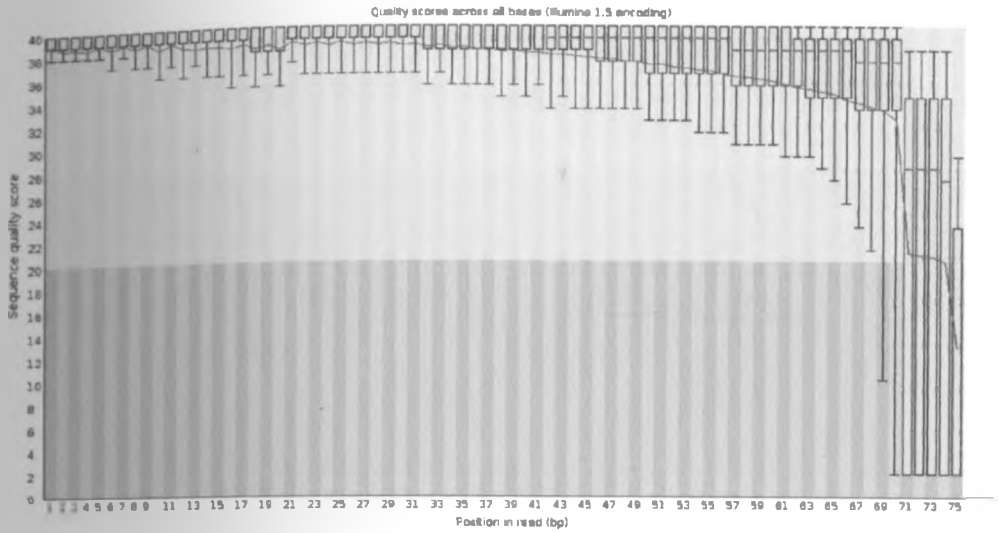


Figure 4: Sequence quality of aligned reads: the x-axis represents positions on the read, y-axis shows different sequence quality scores. and the red lines in the yellow box-plots are the means of quality score. The green, brown and pink coloured segments represent good, moderately-good, and poor sequence quality regions.

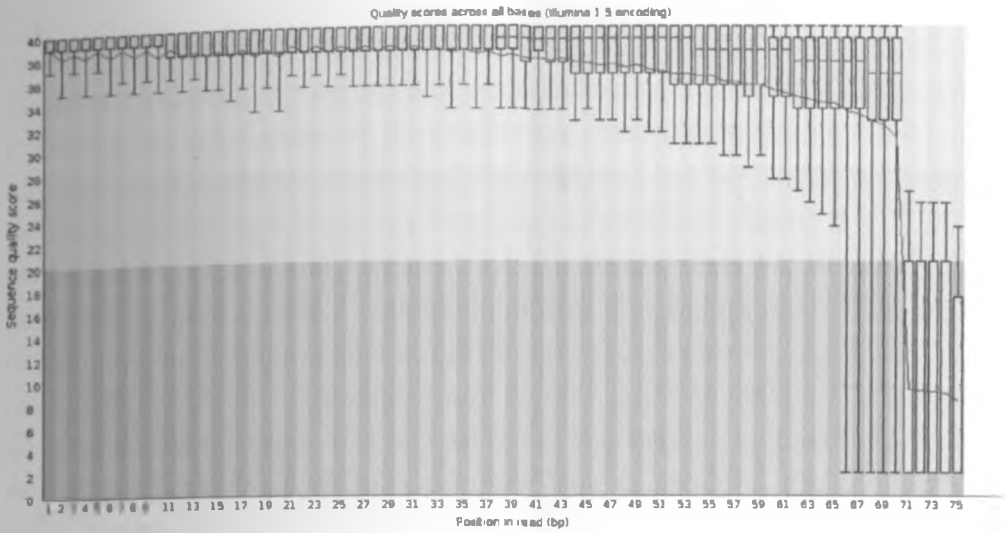


Figure 5: Sequence quality of unaligned reads: the x-axis represents positions on the read, y-axis shows different sequence quality scores, and the red lines in the yellow box-plots are the means of quality score. The green, brown and pink coloured segments represent good, moderately-good, and poor sequence quality regions.

4.3 Short reads cluster sizes distribution

Clustering of aligned, unaligned and simulated reads resulted in a characteristic distribution of frequency of the sizes of the clusters in which unaligned and simulated reads exhibited many small-sized clusters but few significantly large clusters, a distribution commonly referred to as Power-Law (figures 6 & 7). This distribution was also observed when a subset of aligned reads was clustered. The distribution of read counts overlapping putative TF-binding regions has previously been shown to follow the same distribution in both experimental (Rozowsky *et al.*, 2009) and ChIP-Seq simulated data sets (Zhang *et al.*, 2008). This suggests that some unaligned reads data sets contained potentially legitimate reads with mappable properties, an assertion that was later validated as discussed in the following sections.

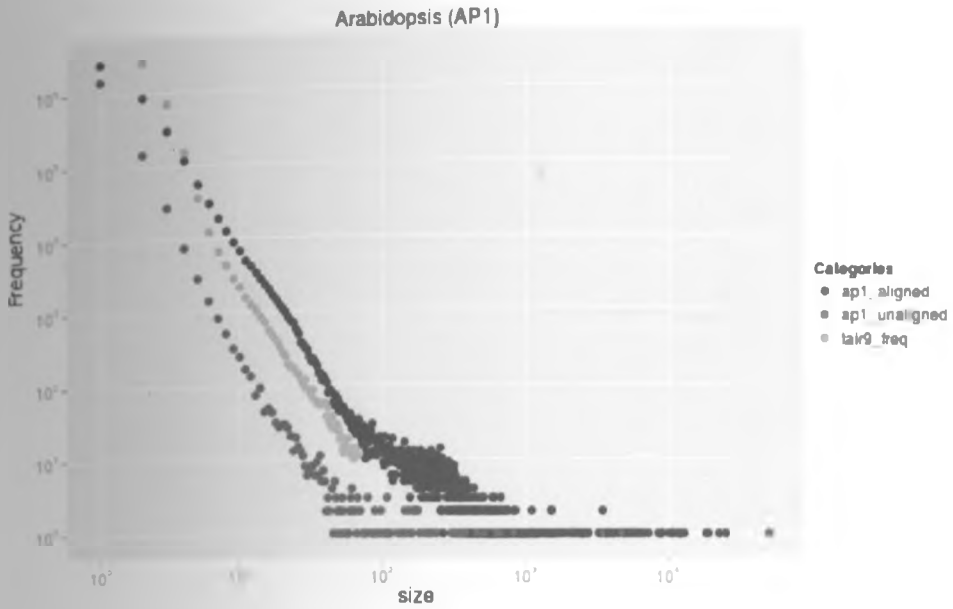


Figure 6: Power-law distribution of frequency of read clusters in an Arabidopsis ChIP-Seq dataset from *Apetala1* (AP1) transcription factor: x-axis represent sizes of clusters and y-axis represent the frequency of occurrence of a particular size of cluster. The distribution is same in aligned (*ap1_aligned*), unaligned (*ap1_unaligned*) and simulated (*tai9_freq*) read clusters.

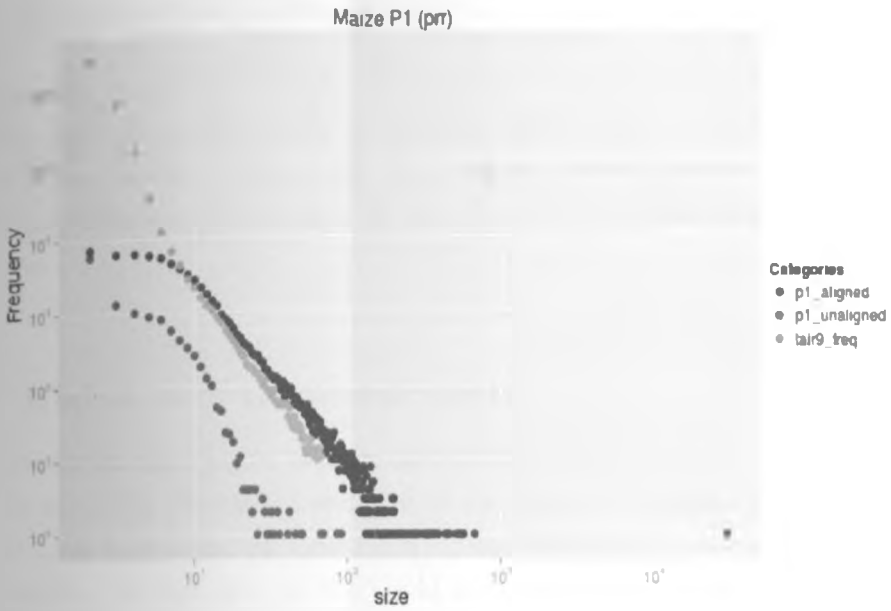


Figure 7: Power-law distribution of frequency of read clusters in a Maize ChIP-Seq dataset from Pericarp1 (P1) transcription factor: x-axis represent sizes of clusters and y-axis represent the frequency of occurrence of a particular size of cluster. The distribution is same in aligned (p1_aligned), unaligned (p1_unaligned) and simulated (tai9_freq) read clusters.

4.4 Taxonomic classification of unaligned reads

We performed a nucleotide database search of cluster seeds using NCBI's BLAST+ stand-alone tool as the first step in determining likely sources of unaligned reads. For each of the BLAST outputs from the unaligned reads data sets, reads were classified into taxonomic units based on their BLAST hits while using NCBI taxonomy tree.

4.4.1 Higher rank taxonomic units

Figure 8 shows the main higher rank taxonomic units represented in the unaligned reads data sets. The aforementioned control data sets of simulated and a subset of aligned reads had almost all their reads classified in the same taxonomic units as their source. For instance, all the reads simulated from Arabidopsis (the six rightmost bars in the Arabidopsis OTUs plot of figure 8 in the topleft panel, represents control datasets of aligned ChIP-Seq and simulated reads) and *C. elegans* (*C. elegans* genome simulated) genomes were assigned to the Plantae and Metazoan taxonomic units respectively. In fact, all the simulated and aligned control reads in these two groups (Arabidopsis and *C. elegans*) were correctly assigned to their respective lower rank taxonomic units, further validating the robustness of the BLAST and taxonomic assignment processes employed in the study.

Proportions of unaligned reads assigned to each of the higher rank taxonomic units varied considerably across different data sets and genomes. Human data sets (figure 9) for instance exhibited varying proportions of metazoan and bacterial sequences, and we observed some ChIP-Seq runs in which a significant amount of unaligned reads were assigned into the metazoan group. In one such set of human ChIP-Seq runs (SRR054870, SRR054871, SRR054881, SRR054882, SRR054883, SRR054884, SRR054885, SRR054892, SRR054894, and SRR054895) generated from an ENCODE study that was aimed at determining the DNA-binding profiles of human Erythroblast Transformation Specific (ETS) family of transcription factors ERG, ELF1, SPI1 and SPDEF (Wei *et al.*, 2010), 80% of the unaligned reads in each run were assigned to the metazoan taxonomic unit, suggesting the

presence of potentially legitimate human-derived reads. Indeed a much closer examination involving assignment of reads into lower rank taxonomic units revealed the presence of significant amounts of human reads in these data sets (especially in SRR054882, SRR054883 and SRR070251), in which more than half of the total number of unaligned reads were classified as sequences derived from the human genome (figures 10 & 11).

The presence of a significant amount of metazoan reads was also observed in the other metazoan genome data sets of *Drosophila* and *C. elegans*, in which some *Drosophila* (SRR067915, SRR086223) and *C. elegans* ChIP-Seq data sets (SRR0107326) had a significant proportion of reads assigned to *D. melanogaster* (figure 11a) and *C. elegans* taxonomic units respectively (figure 11b). Additionally, metazoan sequences were found in both *Arabidopsis* and maize data sets (figure 8), although the latter had relatively less proportions of unaligned reads in the metazoan taxonomic unit. These metazoan-derived sequences were later shown to be mainly contaminant sequences as reported below.

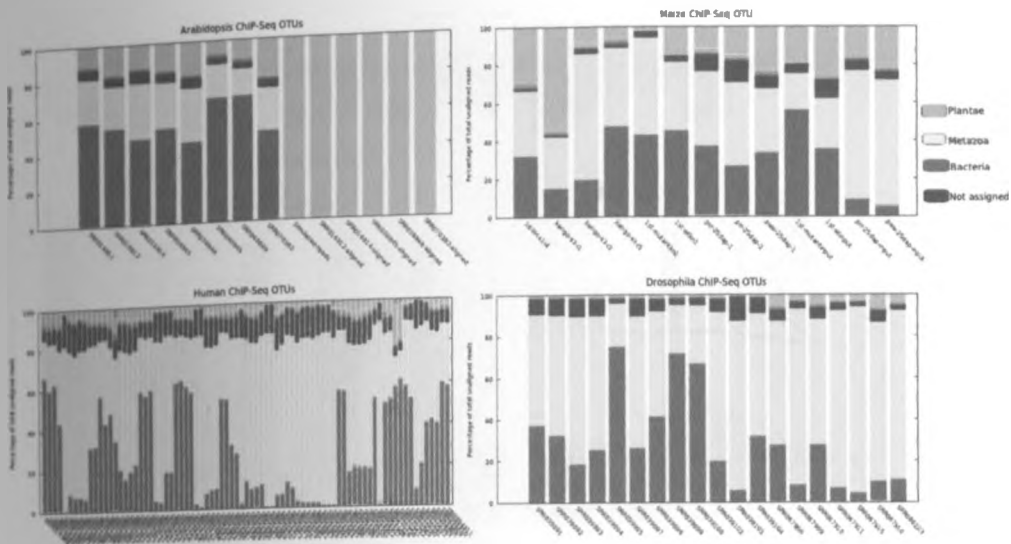


Figure 8: Taxonomic classification of unaligned reads: each bar on the x-axis represents a ChIP-Seq run, and the y-axis shows the percentage of the total unaligned reads. Bacteria and Metazoa are the major higher rank OTUs. Reads simulated from Arabidopsis (the six rightmost bars in the topleft panel, represents control datasets of aligned ChIP-Seq and simulated reads) and *C. elegans* (*C. elegans* genome simulated) genomes were assigned to the Plantae and Metazoan taxonomic units, respectively.

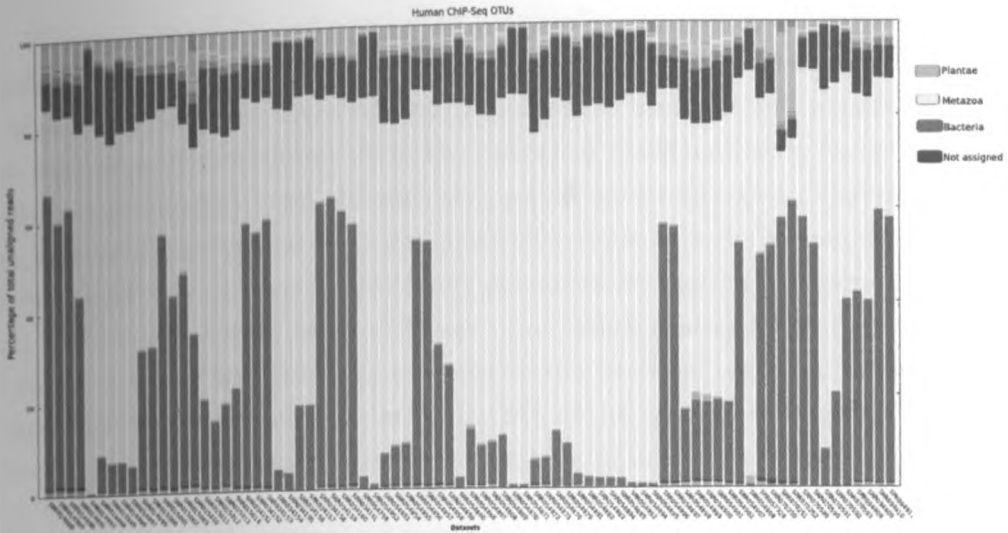


Figure 9: Taxonomic classification of Human ChIP-Seq unaligned reads: each bar on the x-axis represent a ChIP-Seq run, y-axis shows percentage of total unaligned reads. High proportions of bacterial and metazoan sequences are observed.

Whereas several data sets contained potentially legitimate reads, contaminant sequences were found in appreciable amounts in unaligned reads data sets of the five genomes analyzed. Accordingly, sequences of bacterial and metazoan origin were identified in almost all of the unaligned reads data sets, except in the control data sets (figure 8), although the proportion varied across different ChIP-Seq runs and experiments. Two particular experiments in the human ENCODE project that were aimed at identifying genome-wide binding patterns of PolII and TAL1 transcription factors (Birney *et al.*, 2007) had considerable amounts of reads in their ChIP-Seq runs assigned to the bacterial taxonomic unit, with some ChIP-Seq runs having as much as 60% of the unaligned reads being classified as bacterial sequences. In depth analysis of one of these data sets from the TAL1 ChIP-Seq experiment (SRR70589) revealed the presence of predominantly three different species of bacteria: *Escherichia coli*, *Propionibacterium acnes* and Enterobacteriaceae (figure 8). Additionally, mice sequences were identified in two (SRR070251, SRR070251) of the TAL1 ChIP-Seq data sets (figure 8). Relatively higher proportions of bacterial contamination from *Meiothermus silvanus* were also observed in Drosophila HSF binding data sets (SRR039095, SRR039099, SRR039100), and Enterobacteriaceae in *C. elegans* DAF-16 data sets (SRR017602, SRR017605) contained significant amounts of both Enterobacteriaceae and *Escherichia coli* sequences (figure 10).

Interestingly, human sequences were the main source of metazoan contamination for most of the data sets analysed. At least one data set of unaligned reads in each of *D. melanogaster*, *C. elegans*, *A. thaliana* and maize ChIP-Seq experiments contained considerable amounts of human sequences, as depicted in figure 10. In a *Drosophila* ChIP-Seq study aimed at examining genome-wide distribution of *Drosophila* Heat Shock Factor (HSF) binding (Guertin and Lis, 2010), several ChIP-Seq runs (SRR039095, SRR039098, SRR039099, SRR039100, SRR039103) contained significant amounts of human sequences (figure 10). In addition, some runs from modENCODE projects for identification of *C. elegans* DAF-16 (SRR017601, SRR017603, SRR017604) and LIN-11 (SRR107307, SRR107355) TF binding sites contained human sequences, as well as Arabidopsis SEP3 (Kaufmann *et al.*, 2009) (SRR016811, SRR016812) and AP1 (Kaufmann *et al.*, 2010b) (SRR038848) binding ChIP-Seq data sets (figure 10). The maize P1 TF binding (Morohashi *et al.*, 2012) data sets also contained human contaminant sequences, as well as contaminant sequences from *Salmo salar* sequences, whose most likely source would be salmon sperm DNA (figure 10). Contamination with salmon sperm DNA can be easily explained by the experimental procedures used that involve a pre-clearing step with covalently linked salmon sperm DNA to Protein A beads (Morohashi *et al.*, 2012, 2009).

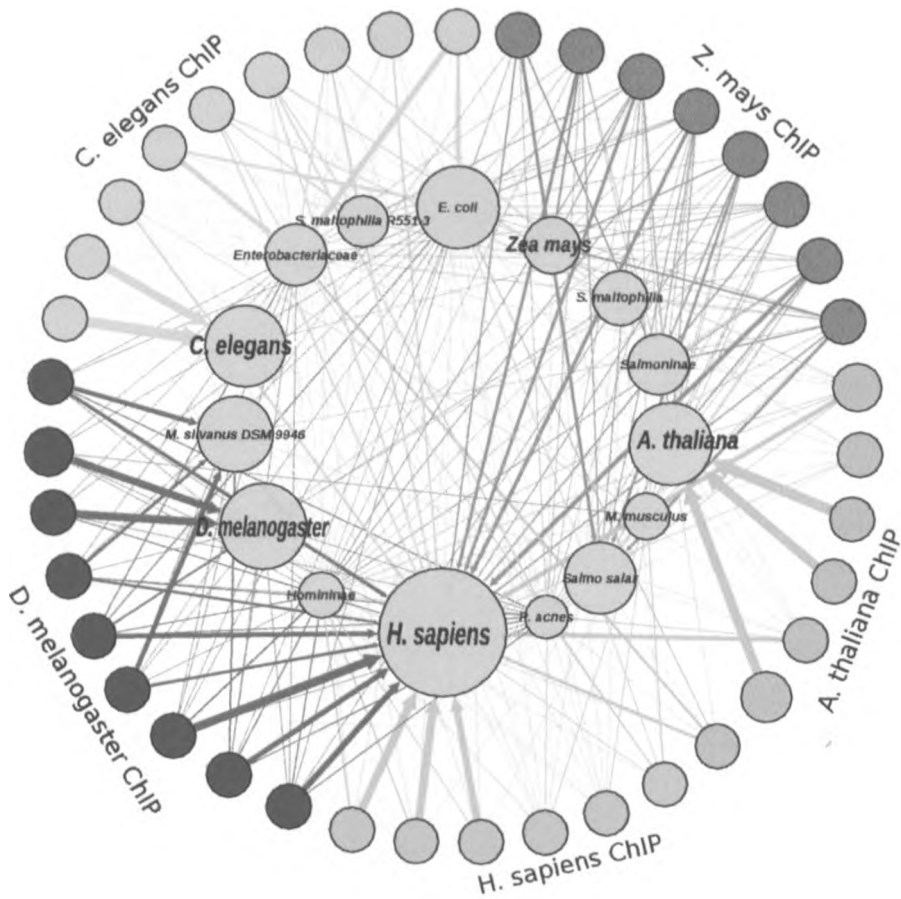
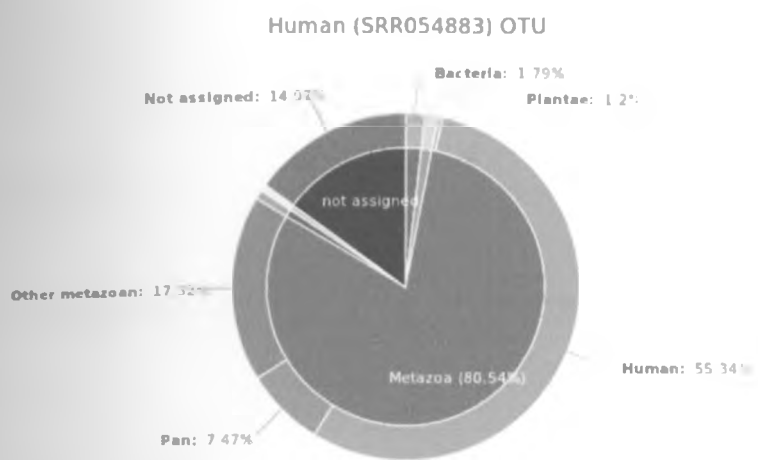
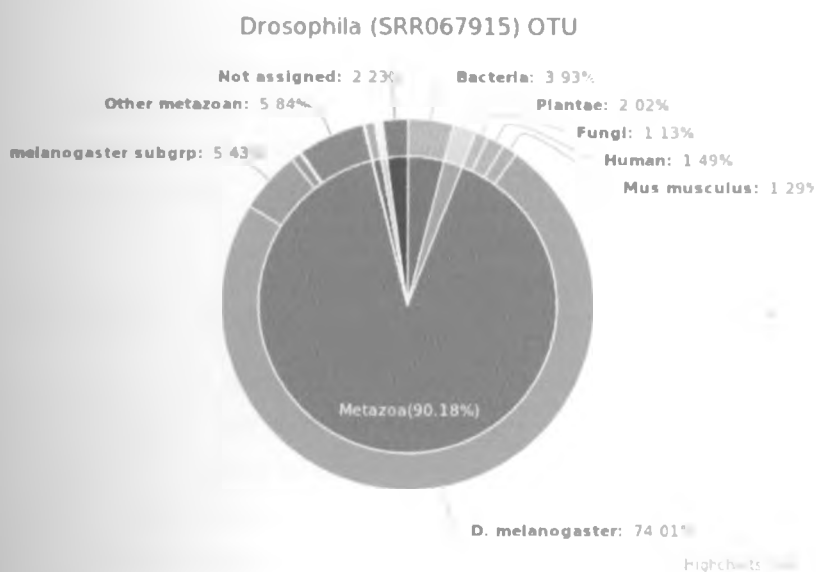


Figure 10: Relative abundance of taxonomic units in ChIP-Seq data sets: outer circles represent different ChIP-Seq runs colour-coded to represent different genomes; inner circles represent taxonomic units with sizes proportional to relative abundance. Arrow sizes are proportional to the amount of reads assigned into the taxonomic unit(s) they are pointing to. Lower rank OTU assignment reveals presence of both potentially legitimate reads as well as contaminant sequences.



(a) Human ChIP-Seq reads OTU



(b) Drosophila ChIP-Seq reads OTU

Figure 11: Taxonomic classification of selected Human (a) and Drosophila (b) unaligned reads: more than half of the unaligned reads in these data sets are potentially legitimate reads.

Chapter 5

Discussion and Conclusion

5.1 Discussion

This study- which involved a series of analyses of short sequence reads from NCBI SRA generated from ChIP-Seq experiments aimed at determining TFBSs in well-studied organisms, reveals a twofold origin of unaligned reads in these experiments: (1) genome-derived potentially legitimate reads that fail to align to the reference genome due to genome-related complexities or alignment program biases; and (2) sequences that arise due to possible contamination of ChIP-Seq libraries with foreign DNA.

5.1.1 Variations in alignment proportions

Comparisons of alignment proportions across different ChIP-Seq runs in different experiments indicate that while there may be standardized protocols in the ChIP-Seq process in individual laboratories, variations in alignment proportions are still observed even within experiments from the same laboratory, with even more variation in different genomes. For instance, both the ENCODE and mod-ENCODE projects have such elaborate and standardized guidelines for ChIP-Seq experiments in order to generate high quality and comparable data, but reported here are variations in alignment proportions in different ChIP-Seq runs and experiments within these two major projects. This is an indication that generally, failure of a subset of reads to align to the reference genome cannot be entirely

attributed to experiment-specific protocols since despite the same protocol being used in ChIP-Seq experiments, large variations in proportions of aligned reads are observed in individual ChIP-Seq runs. Additionally, the generation of more sequences in a ChIP-Seq experiment does not influence alignment success. These observations indicate that failure of reads to align to the reference genome is not entirely laboratory specific, instead it suggests other causal factors.

Whereas the presence of sequences with poor quality did not significantly affect alignment proportions in this study, we would like to point out that the effect could be more pronounced in cases where these sequences form a significant amount of the short read data sets, that is, the larger the amount of poor quality sequences the higher the likelihood of reads not aligning to the reference genome (Kaufmann *et al.*, 2010a). Ensuring good quality ChIP-Seq reads is largely dependent on the sequencing libraries preparation step, and even after sequencing, care must be taken to remove low quality reads in the preprocessing step of the ChIP-Seq data analysis pipeline. Moreover, since ChIP-Seq analysis involves the use of uniquely-aligned reads in identification of TFBSs, there is a missed opportunity in identification of TFBSs in highly repetitive genome sites, particularly in human and maize genomes (Kronmiller and Wise, 2009; SanMiguel *et al.*, 1998; Schnable *et al.*, 2009), when multi-reads are discarded. This can result in potential false negatives (Blahnik *et al.*, 2010). Nonetheless, the development of ChIP-Seq analysis algorithms for the discovery of TFBSs in highly repetitive genomes is an active area of research (Chung *et al.*, 2011; Wang *et al.*, 2010).

Genetic variation within species could also be a contributing factor in the observed unaligned reads, since extensive genomic variation and diversity has been described in eukaryotes, especially in human (Barbujani *et al.*, 1997; Bennett *et al.*, 2004; Durbin *et al.*, 2010; Ewing and Kazazian, 2010; Iskow *et al.*, 2010; Xing *et al.*, 2009) and maize genomes (Dooner and He, 2008; Haberer *et al.*, 2005; Schnable *et al.*, 2009), in which most of the diversity is accounted for by transposable elements comprising nearly half of the human genome (Lander *et al.*, 2001) and an estimated 50 to 80% of the maize (Meyers *et al.*, 2001) genome. Li R. (Li *et al.*, 2010) and colleagues have shown that integrating an Asian and an African human genome with the NCBI human reference genome results in identification of approximately 5 Mb of novel sequences not present in the ref-

erence genome. In their effort to construct a complete human-pan genome, it was estimated that the complete genome would contain about 20-40 Mb of novel sequences. Clearly, there is a loss of information when reads are aligned to a single genome. However, large-scale genome resequencing projects like the human 1000 genomes project (<http://www.1000genomes.org/>), *A. thaliana* 1001 genomes project (<http://1001genomes.org>) and the Drosophila population genomics project (<http://dpgp.org>) offer an excellent opportunity for the alignment of reads to multiple genomes thereby significantly reducing the loss of information.

5.1.2 Potentially legitimate reads

Failure to align potentially legitimate reads to their reference genome could also be attributed to the manner in which alignment algorithms align short reads derived from next generation sequencing technologies. Technically, these next generation sequence alignment programs, unlike BLAST, do map reads heuristically

To substantiate our assertion that these potentially legitimate reads are true reads with mappable properties, we successfully realigned between a half and two thirds of the potentially legitimate reads back to their respective reference genomes using SHRiMP (Rumble *et al.*, 2009), an alignment program with capabilities of not only mapping short reads in highly polymorphic regions of the genome, but also of handling indels and allowance of mismatches. There seems to be a cost of loss of sensitivity when using more memory and time efficient alignment algorithms. The caveat, however, is that SHRiMP does not seem to significantly outperform bowtie in the initial alignment process. In fact, there is no gold standard short read alignment algorithm (Schbath *et al.*, 2012). Nevertheless, we would like to point out that that SHRiMP's apparent superiority and ability to realign a considerable amount of reads to the reference genome is largely due to the fact that it is realigning a subset of reads that have been shown to be potentially legitimate by our analyses. The rationale behind this is that since short read alignment algorithms map reads in a heuristic manner, there is increased likelihood of correctly mapping more legitimate reads when the set of reads to be mapped has been filtered and hence contains significantly

high proportion of legitimate reads with increased likelihood of mapping to the reference genome heuristically. The success of alignment is therefore augmented when an alignment algorithm has capabilities of both gapped-alignment and ability to deal with indels in reads, as is the case with SHRiMP. It should be noted, however, that the analyses reported here concentrated on determining the origin of unaligned reads rather than a detailed evaluation of performance of short read alignment algorithms, which has been reported elsewhere (Ruffalo *et al.*, 2011; Schbath *et al.*, 2012).

5.1.3 Contamination

Taxonomic classification was also able to uncover foreign DNA in the ChIP-Seq data sets. Although not extensively reported in literature, it is not uncommon to identify contaminant sequences in ChIP-Seq data sets (Kaufmann *et al.*, 2010a). One possible source of contamination could be adapter sequences, which are ligated to the DNA fragments before sequencing on an Illumina platform. The proportion of adapter sequences usually varies across different sequencing libraries, and is dependent on the quality of the library (Kaufmann *et al.*, 2010a). However, we observed low proportions of adapter sequences that have an insignificant effect on alignment of reads. One main source of contamination in at least one of the data sets from the genomes analysed was human sequences. This could presumably be due to contamination from researchers performing the experiment and/or extensive contamination of public sequence databases with human DNA, as shown by an in depth sequence search in NCBI, Ensembl, JGI, and UCSC carried out by Longo and colleagues (Longo *et al.*, 2011).

On the other hand, it is now widely accepted that the human genome contains bacterial footprints, with the International human Genome Sequencing Consortium reporting in early 2001 that between 113 and 223 human genes are of bacterial origin, although Salzberg and colleagues (Salzberg *et al.*, 2001) obtained a lower gene count upon recalculation. Nonetheless, higher proportions of unaligned reads assigned to bacterial genomes reflects either the presence of bacterial sequences in the human genome or sequencing of the human microbiome

(Gill *et al.*, 2006), or both. Additionally, bacterial sequences in the unaligned reads data sets could arise due to sequencing of contaminant bacteria that are abundant in molecular biology laboratories, especially mycoplasma bacteria. The presence of bacterial sequences in other data sets from the other genomes analysed is largely due to contamination of the sequencing libraries with bacteria. This could be the case in the observed high amounts of *E. coli* and other Enterobacterial sequences in selected *C. elegans* ChIP-Seq runs (figure 10).

5.2 Conclusion

While failure of ChIP-Seq reads to align back to the reference genome is not uncommon, this phenomenon has never been investigated in detail. We have shown that this failure is not dependent on sequence quality, the amount of sequences generated in a ChIP-Seq experiment, or the laboratory performing the experiments; rather it is mainly due to contamination of ChIP-Seq sequencing libraries with foreign DNA material. The main source of contamination in most experiments was shown to be bacteria and metazoa of which human sequences was dominant. *E. coli*, *Meiothermus silvanus* and Enterobacteria contamination was predominantly in *C. elegans* and *D. melanogaster* ChIP-Seq studies. Interestingly, contamination from the fish *Salmo salar* DNA was observed in some maize ChIP-Seq data sets. This contaminant, as mentioned earlier, is likely to have been due to the use of Salmon sperm as a blocking agent that prevents non-specific binding in the ChIP-Seq process. In addition, this study has revealed the presence of potentially legitimate reads in selected unaligned ChIP-Seq data sets, implying that the choice of the alignment algorithm contributes to the efficiency of the alignment process.

We strongly recommend that researchers performing ChIP-Seq experiments take utmost care while preparing their sequencing libraries since it is mainly at this stage that foreign DNA materials are introduced into the sample. We also recommend an exploration of unaligned reads since it has been shown that these data sets may contain legitimate reads for ChIP-Seq analysis.

Research on improving the sensitivity of short read alignment algorithms is

also highly recommended.

Appendix 1

Human ChIP-Seq data sets:

SRR014988: STAT1 Transcription Factor in Human HeLa S3 Anti-STAT1 Ab
SRR014989: STAT1 Transcription Factor in Human HeLa S3 Anti-STAT1 Ab
SRR014990: STAT1 Transcription Factor in Human HeLa S3 Anti-STAT1 Ab
SRR014991: STAT1 Transcription Factor in Human HeLa S3 Anti-STAT1 Ab
SRR014993: STAT1 Transcription Factor in Human HeLa S3 Input DNA
SRR014994: STAT1 Transcription Factor in Human HeLa S3 Input DNA
SRR014995: STAT1 Transcription Factor in Human HeLa S3 Input DNA
SRR014996: STAT1 Transcription Factor in Human HeLa S3 Input DNA
SRR014997: STAT1 Transcription Factor in Human HeLa S3 Input DNA
SRR014999: PolII Transcription Factor in Human HeLa S3 Mouse mAb
SRR015000: PolII Transcription Factor in Human HeLa S3 Mouse mAb
SRR015001: PolII Transcription Factor in Human HeLa S3 Mouse mAb
SRR015002: PolII Transcription Factor in Human HeLa S3 Mouse mAb
SRR015003: PolII Transcription Factor in Human HeLa S3 Mouse mAb
SRR015010: PolII Transcription Factor in Human HeLa S3 Input DNA
SRR015011: PolII Transcription Factor in Human HeLa S3 Input DNA
SRR015012: PolII Transcription Factor in Human HeLa S3 Input DNA
SRR015013: PolII Transcription Factor in Human HeLa S3 Input DNA
SRR015014: PolII Transcription Factor in Human HeLa S3 Input DNA
SRR034151: ETS1 and RUNX binding pETS1 Ab
SRR034152: ETS1 and RUNX binding pETS1 Ab
SRR034153: ETS1 and RUNX binding pETS1 Ab
SRR034154: ETS1 and RUNX binding mRUNX Ab
SRR034155: ETS1 and RUNX binding mRUNX Ab

SRR034156: ETS1 and RUNX binding pCBP Ab
SRR034157: ETS1 and RUNX binding pCBP Ab
SRR034158: ETS1 and RUNX binding Input DNA
SRR034159: ETS1 and RUNX binding Input DNA
SRR034160: ETS1 and RUNX binding Input DNA
SRR034161: ETS1 and RUNX binding Input DNA
SRR054758: Human ETS family of Tfs TF specific Ab
SRR054852: Human ETS family of Tfs TF specific Ab
SRR054853: Human ETS family of Tfs TF specific Ab
SRR054854: Human ETS family of Tfs TF specific Ab
SRR054855: Human ETS family of Tfs TF specific Ab
SRR054856: Human ETS family of Tfs IgG
SRR054857: Human ETS family of Tfs IgG
SRR054858: Human ETS family of Tfs TF specific Ab
SRR054859: Human ETS family of Tfs TF specific Ab
SRR054860: Human ETS family of Tfs IgG
SRR054866: Human ETS family of Tfs TF specific Ab
SRR054867: Human ETS family of Tfs TF specific Ab
SRR054868: Human ETS family of Tfs TF specific Ab
SRR054869: Human ETS family of Tfs TF specific Ab
SRR054870: Human ETS family of Tfs TF specific Ab
SRR054871: Human ETS family of Tfs IgG
SRR054872: Human ETS family of Tfs IgG
SRR054873: Human ETS family of Tfs IgG
SRR054876: Human ETS family of Tfs TF specific Ab
SRR054879: Human ETS family of Tfs IgG
SRR054881: Human ETS family of Tfs TF specific Ab
SRR054882: Human ETS family of Tfs TF specific Ab
SRR054883: Human ETS family of Tfs TF specific Ab
SRR054884: Human ETS family of Tfs TF specific Ab
SRR054885: Human ETS family of Tfs TF specific Ab
SRR054892: Human ETS family of Tfs TF specific Ab
SRR054894: Human ETS family of Tfs TF specific Ab

SRR054895: Human ETS family of Tfs TF specific Ab
SRR054896: Human ETS family of Tfs IgG
SRR054897: Human ETS family of Tfs IgG
SRR054898: Human ETS family of Tfs TF specific Ab
SRR054899: Human ETS family of Tfs TF specific Ab
SRR054900: Human ETS family of Tfs TF specific Ab
SRR054901: Human ETS family of Tfs TF specific Ab
SRR054902: Human ETS family of Tfs TF specific Ab
SRR054907: Human ETS family of Tfs IgG
SRR054908: Human ETS family of Tfs Input DNA*
SRR057328: Human ETS family of Tfs IgG
SRR070250: GABP-alpha binding GABP-alpha Ab
SRR070251: GABP-alpha binding IgG
SRR070252: GABP-alpha binding IgG
SRR070589: TAL1 in hematopoietic lineages TF specific Ab
SRR070590: TAL1 in hematopoietic lineages TF specific Ab
SRR070591: TAL1 in hematopoietic lineages TF specific Ab
SRR070592: TAL1 in hematopoietic lineages TF specific Ab
SRR070593: TAL1 in hematopoietic lineages TF specific Ab
SRR094805: ChIP-Seq in expanded hematopoietic and progenitor cells WCE*
SRR094806: ChIP-Seq in expanded hematopoietic and progenitor cells TF specific
SRR094807: ChIP-Seq in expanded hematopoietic and progenitor cells TF specific
SRR094808: ChIP-Seq in expanded hematopoietic and progenitor cells TF specific
SRR094809: ChIP-Seq in expanded hematopoietic and progenitor cells TF specific
SRR094810: ChIP-Seq in expanded hematopoietic and progenitor cells TF specific
SRR094811: ChIP-Seq in expanded hematopoietic and progenitor cells TF specific
SRR094812: ChIP-Seq in expanded hematopoietic and progenitor cells TF specific
SRR094813: ChIP-Seq in expanded hematopoietic and progenitor cells TF specific
SRR094814: ChIP-Seq in expanded hematopoietic and progenitor cells TF specific

Drosophila ChIP-Seq data sets:

SRR039091: S2_20HS_PREIMMUNE

SRR039092: S2_20HS_PREIMMUNE

SRR039093: S2_20HS_HSF_IP

SRR039094: S2_20HS_HSF_IP

SRR039095: S2_20HS_HSF_IP

SRR039096: S2_20HS_HSF_KD-IP

SRR039097: S2_20HS_HSF_KD-IP

SRR039098: S2_20HS_HSF_KD-IP

SRR039099: S2_NHS_HSF_IP

SRR039100: S2_NHS_HSF_IP

SRR039101: S2_NHS_HSF_IP

SRR039102: S2_NHS_HSF_IP

SRR039103: S2_NHS_HSF_KD-IP

SRR039104: S2_NHS_HSF_KD-IP

SRR067906: S2-DRSC-ChIP-Pc

SRR067909: S2-DRSC-ChIP-Ph

SRR067910: S2-DRSC-ChIP-Psc

SRR067915: S2-DRSC-ChIP-Input

SRR067916: S2-DRSC-ChIP-H3K4me3

C. elegans ChIP-Seq data sets:

SRR017601 Young adult replicate 1 POL II
SRR017602 Young adult replicate 1 Input
SRR017603 Young adult replicate 2 GFP
SRR017604 Young adult replicate 2 POL II
SRR017605 Young adult replicate 2 Input
SRR107302 Snyder-LIN-11-GFP-L2-rep1 extraction1-seq1 aliquote 1
SRR107305 Snyder-LIN-11-GFP-L2-rep1 extraction1-seq1 aliquote 2
SRR107306 Snyder-LIN-11-GFP-L2-rep2 extraction2-seq1 aliquote 1
SRR107307 Snyder-LIN-11-GFP-L2-rep2 extraction2-seq1 aliquote 2
SRR107308 Snyder-UNC-130-GFP-L1-rep1 extraction1-seq1 aliquote 1
SRR107309 Snyder-UNC-130-GFP-L1-rep1 extraction1-seq1 aliquote 2
SRR107310 Snyder-UNC-130-GFP-L1-rep2 extraction2-seq1 aliquote 1
SRR107311 Snyder-UNC-130-GFP-L1-rep2 extraction2-seq1 aliquote 2
SRR107312 Snyder-HLH-1-GFP-emb-rep1 extraction1-seq1 aliquote 1
SRR107313 Snyder-HLH-1-GFP-emb-rep1 extraction1-seq1 aliquote 2
SRR107314 Snyder-HLH-1-GFP-emb-rep2 extraction2-seq1 aliquote 1
SRR107315 Snyder-HLH-1-GFP-emb-rep2 extraction2-seq1 aliquote 2
SRR107317 Snyder-NHR-6-GFP-L2-rep1 extraction1-seq1 aliquote 1
SRR107319 Snyder-NHR-6-GFP-L2-rep1 extraction1-seq1 aliquote 2
SRR107320 Snyder-NHR-6-GFP-L2-rep2 extraction2-seq1 aliquote 1
SRR107322 Snyder-NHR-6-GFP-L2-rep2 extraction2-seq1 aliquote 2
SRR107324 Snyder-N2-POLII-eemb-rep1 extraction1-seq1 aliquote 1
SRR107325 Snyder-N2-POLII-eemb-rep1 extraction1-seq1 aliquote 2
SRR107326 Snyder-N2-POLII-eemb-rep2 extraction2-seq1 aliquote 1
SRR107327 Snyder-N2-POLII-eemb-rep2 extraction2-seq1 aliquote 2
SRR107329 Snyder-N2-POLII-lemb-rep1 extraction1-seq1 aliquote 1
SRR107330 Snyder-N2-POLII-lemb-rep1 extraction1-seq1 aliquote 2
SRR107331 Snyder-N2-POLII-lemb-rep2 extraction2-seq1 aliquote 1
SRR107332 Snyder-N2-POLII-lemb-rep2 extraction2-seq1 aliquote 2
SRR107339 Snyder-N2-POLII-L1-rep1 extraction1-seq1 aliquote 1
SRR107340 Snyder-N2-POLII-L1-rep1 extraction1-seq1 aliquote 2
SRR107341 Snyder-N2-POLII-L1-rep2 extraction2-seq1 aliquote 1

SRR107342 Snyder-N2-POLII-L1-rep2 extraction2-seq1 aliquote 2
SRR107345 Snyder-N2-POLII-L2-rep1 extraction1-seq1 aliquote 1
SRR107348 Snyder-N2-POLII-L2-rep1 extraction1-seq1 aliquote 2
SRR107349 Snyder-N2-POLII-L2-rep2 extraction2-seq1 aliquote 1
SRR107350 Snyder-N2-POLII-L2-rep2 extraction2-seq1 aliquote 2
SRR107352 Snyder-N2-POLII-L3-rep1 extraction1-seq1 aliquote 1
SRR107355 Snyder-N2-POLII-L3-rep1 extraction1-seq1 aliquote 2.
SRR107356 Snyder-N2-POLII-L3-rep2 extraction2-seq1 aliquote 1
SRR107360 Snyder-N2-POLII-L3-rep2 extraction2-seq1 aliquote 2
SRR107536 Snyder-N2-POLII-YA-rep1 extraction1-seq1 aliquote 1
SRR107537 Snyder-N2-POLII-YA-rep1 extraction1-seq1 aliquote 2
SRR107538 Snyder-N2-POLII-YA-rep2 extraction2-seq1 aliquote 1
SRR107539 Snyder-N2-POLII-YA-rep2 extraction2-seq1 aliquote 2
SRR107540 Snyder-N2-POLII-L4-rep1 extraction1-seq1 aliquote 1
SRR107543 Snyder-N2-POLII-L4-rep1 extraction1-seq1 aliquote 2
SRR107544 Snyder-N2-POLII-L4-rep2 extraction2-seq1 aliquote 1
SRR107545 Snyder-N2-POLII-L4-rep2 extraction2-seq1 aliquote 2
SRR107546 Snyder-GEI11-GFP-L4-rep1 extraction1-seq1 aliquote 1
SRR107547 Snyder-GEI11-GFP-L4-rep1 extraction1-seq1 aliquote 2
SRR107548 Snyder-GEI11-GFP-L4-rep2 extraction2-seq1 aliquote, 1
SRR107549 Snyder-GEI11-GFP-L4-rep2 extraction2-seq1 aliquote 2
SRR107550 Snyder-PHA4-GFP-lemb-rep1 extraction1-seq1 aliquote 1
SRR107551 Snyder-PHA4-GFP-lemb-rep1 extraction1-seq1 aliquote 2
SRR107552 Snyder-PHA4-GFP-lemb-rep2 extraction2-seq1 aliquote 1
SRR107553 Snyder-PHA4-GFP-lemb-rep2 extraction2-seq1 aliquote 2
SRR107554 Snyder-MEP-1-GFP-emb-rep1 extraction1-seq1 aliquote 1
SRR107555 Snyder-MEP-1-GFP-emb-rep1 extraction1-seq1 aliquote 2
SRR107556 Snyder-MEP-1-GFP-emb-rep2 extraction2-seq1 aliquote 1
SRR107557 Snyder-MEP-1-GFP-emb-rep2 extraction2-seq1 aliquote 2
SRR107558 Snyder-MDL-1-GFP-L1-rep1 extraction1-seq1 aliquote 1
SRR107559 Snyder-MDL-1-GFP-L1-rep1 extraction1-seq1 aliquote 2
SRR107560 Snyder-MDL-1-GFP-L1-rep2 extraction2-seq1 aliquote 1
SRR107562 Snyder-LIN-15B-GFP-L3-rep1 extraction1-seq1 aliquote 1

SRR107563 Snyder-LIN-15B-GFP-L3-rep1 extraction1-seq1 aliquote 2
SRR107564 Snyder-LIN-15B-GFP-L3-rep2 extraction2-seq1 aliquote 1
SRR107565 Snyder-LIN-15B-GFP-L3-rep2 extraction2-seq1 aliquote 2
SRR107566 Snyder-BLMP-1-GFP-L1-rep1 extraction1-seq1 aliquote 1
SRR107568 Snyder-BLMP-1-GFP-L1-rep1 extraction1-seq1 aliquote 2
SRR107570 Snyder-BLMP-1-GFP-L1-rep2 extraction2-seq1 aliquote 1
SRR107572 Snyder-BLMP-1-GFP-L1-rep2 extraction2-seq1 aliquote 2
SRR107574 Snyder-LIN-13-GFP-emb-rep1 extraction1-seq1 aliquote 1
SRR107575 Snyder-LIN-13-GFP-emb-rep1 extraction1-seq1 aliquote 2
SRR107576 Snyder-LIN-13-GFP-emb-rep2 extraction2-seq1 aliquote 1
SRR107580 Snyder-ELT-3-GFP-L1-rep1 extraction1-seq1 aliquote 1
SRR107581 Snyder-ELT-3-GFP-L1-rep1 extraction1-seq1 aliquote 2
SRR107583 Snyder-ELT-3-GFP-L1-rep2 extraction2-seq1 aliquote 1
SRR107584 Snyder-ELT-3-GFP-L1-rep2 extraction2-seq1 aliquote 2
SRR107585 Snyder-CEH-30-GFP-lemb-rep1 extraction1-seq1 aliquote 1
SRR107587 Snyder-CEH-30-GFP-lemb-rep2 extraction2-seq1 aliquote 1
SRR107588 Snyder-CEH-30-GFP-lemb-rep2 extraction2-seq1 aliquote 2
SRR107589 Snyder-EGL-27-GFP-L1-rep1 extraction1-seq1 aliquote 1
SRR107590 Snyder-EGL-27-GFP-L1-rep1 extraction1-seq1 aliquote 2
SRR107591 Snyder-EGL-27-GFP-L1-rep2 extraction2-seq1 aliquote 1
SRR107592 Snyder-EGL-27-GFP-L1-rep2 extraction2-seq1 aliquote 2
SRR107593 Snyder-SKN-1-GFP-L1-rep1 extraction1-seq1 aliquote 1
SRR107594 Snyder-SKN-1-GFP-L1-rep1 extraction1-seq1 aliquote 2
SRR107595 Snyder-SKN-1-GFP-L1-rep2 extraction2-seq1 aliquote 1
SRR107596 Snyder-SKN-1-GFP-L1-rep2 extraction2-seq1 aliquote 2
SRR107597 Snyder-PQM-1-GFP-L3-rep1 extraction1-seq1 aliquote 1
SRR107598 Snyder-PQM-1-GFP-L3-rep1 extraction1-seq1 aliquote 2
SRR107600 Snyder-PQM-1-GFP-L3-rep2 extraction2-seq1 aliquote 1
SRR107601 Snyder-PQM-1-GFP-L3-rep2 extraction2-seq1 aliquote 2
SRR107603 Snyder-LIN-39-GFP-L3-rep1 extraction1-seq1 aliquote 1
SRR107604 Snyder-LIN-39-GFP-L3-rep1 extraction1-seq1 aliquote 2
SRR107605 Snyder-LIN-39-GFP-L3-rep2 extraction2-seq1 aliquote 1

Arabidopsis ChIP-Seq data sets:

SRR016810 SEP3 ChIPSeq wild-type-replicate 1 Primary and secondary inflorescences of 5-7 weeks old plants wild type

SRR016811 SEP3 ChIP-Seq wild-type replicate 2 Primary and secondary inflorescences of 5-7 weeks old plants wild type

SRR016812 SEP3 ChIP-Seq sep3 mutant replicate 3 Primary and secondary inflorescences of 5-7 weeks old plants mutant sep3

SRR016813 SEP3 ChIP-Seq wild-type replicate 3 Primary and secondary inflorescences of 5-7 weeks old plants wild type

SRR016814 SEP3 ChIP-Seq ag mutant replicate 3 Primary and secondary inflorescences of 5-7 weeks old plants mutant ag

SRR040045 AP2 Chip replicate 1 InflorescenceCol-0

SRR040046 AP2 Chip replicate 1 InflorescenceCol-0

SRR040047 AP2 Chip replicate 1 InflorescenceCol-0

SRR040048 AP2 Chip replicate 2 InflorescenceCol-0

SRR040049 AP2 control replicate 1 InflorescenceCol-0 SALK-071140

SRR040050 AP2 control replicate 1 Inflorescence Col-0 SALK-071140

SRR040051 AP2 control replicate 2 InflorescenceCol-0 SALK-071140

SRR040050 AP2 control replicate 1 Inflorescence Col-0 SALK-071140

SRR040051 AP2 control replicate 2 InflorescenceCol-0 SALK-071140

SRR070382 LFY control replicate 1 Complete seedlings Col-0

SRR070383 LFY control replicate 2 Complete seedlings Col-0

SRR070384 LFY sample/treatment replicate 1 Complete seedlings Col-0

SRR038845 AP1-GR 2h-induced-sample-1 AP1-GR ap1 cal apical meristematic tissue; 2 h induced with DEX

SRR038846 AP1-GR 2h induced sample 2 AP1-GR ap1 cal apical meristematic tissue; 2 h induced with DEX

SRR038847 AP1-GR 2h induced sample 2 AP1-GR ap1 cal apical meristematic tissue; 2 h induced with DEX

SRR038848 AP1-GR uninduced control 1 AP1-GR ap1 cal apical meristematic tissue; uninduced

SRR038849 AP1-GR uninduced control 1 AP1-GR ap1 cal apical meristematic tissue; uninduced

SRR038850 AP1-GR uninduced control 2 AP1-GR ap1 cal apical meristematic tissue; uninduced

SRR038851 AP1-GR uninduced control 2 AP1-GR ap1 cal apical meristematic tissue; uninduced

SRR038848 AP1-GR uninduced control 1 AP1-GR ap1 cal apical meristematic tissue; uninduced

SRR038849 AP1-GR uninduced control 1 AP1-GR ap1 cal apical meristematic tissue; uninduced

SRR038850 AP1-GR uninduced control 2 AP1-GR ap1 cal apical meristematic tissue; uninduced

SRR038851 AP1-GR uninduced control 2 AP1-GR ap1 cal apical meristematic tissue; uninduced

Bibliography

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410. 7
- Barbujani, G., Magagni, A., Minch, E., and Cavalli-Sforza, L. L. (1997). An apportionment of human dna diversity. *Proc Natl Acad Sci U S A*, 94(9):4516–4519. 39
- Barski, A. and Zhao, K. (2009). Genomic location analysis by chip-seq. *J Cell Biochem*, 107(1):11–18. 5
- Bennett, E. A., Coleman, L. E., Tsui, C., Pittard, W. S., and Devine, S. E. (2004). Natural genetic variation caused by transposable elements in humans. *Genetics*, 168(2):933–951. 39
- Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guig, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C. J., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., DENOEUDE, F., REYMOND, A., KAPRANOV, P., ROZOWSKY, J., ZHENG, D., CASTELO, R., FRANKISH, A., HARROW, J., GHOSH, S., SANDELIN, A., HOFACKER, I. L., BAERTSCH, R., KEEFE,

BIBLIOGRAPHY

D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korb, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W.-K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C.-L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S. E., Fu, Y., Green, E. D., Karaz, U., Siepel, A., Taylor, J., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Cooper, G. M., Asimenos, G., Dewey, C. N., Hou, M., Nikolaev, S., Montoya-Burgos, J. I., Lytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N. R., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Srinivasan, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., , N. I. S. C. C. S. P., , B. C. o. M. H. G. S. C., , W. U. G. S. C., , B. I., , C. H. O. R. I., Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C., Kim, J., Bhang, A. A., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W. H., Ng, P., Shahab, A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J. C., Couttet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix, D. A., Euskirchen, G., Hartman, S., Urban, A. E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T. H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M., Haidar, J. N. S., Yu, Y., Ruan, Y., Iyer, V. R., Green, R. D., Wadelius, C., Farnham, P. J., Ren, B., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith,

BIBLIOGRAPHY

- K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Armengol, L., Bird, C. P., de Bakker, P. I. W., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Woodroffe, A., Davydov, E., Dimas, A., Eyraas, E., Hallgrimsdottir, I. B., Huppert, J., Zody, M. C., Abecasis, G. R., Estivill, X., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V. B., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B., and de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447(7146):799-816. 13, 34
- Blahnik, K. R., Dou, L., O'Geen, H., McPhillips, T., Xu, X., Cao, A. R., Iyengar, S., Nicolet, C. M., Ludscher, B., Korf, I., and Farnham, P. J. (2010). Sole-search: an integrated analysis program for peak detection and functional annotation using chip-seq data. *Nucleic Acids Res*, 38(3):e13. 39
- Cheung, M.-S., Down, T. A., Latorre, I., and Ahringer, J. (2011). Systematic bias in high-throughput sequencing data and its correction by beads. *Nucleic Acids Res*, 39(15):e103. 10, 11
- Chung, D., Kuan, P. F., Li, B., Sanalkumar, R., Liang, K., Bresnick, E. H., Dewey, C., and Kele, S. (2011). Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of chip-seq data. *PLoS Comput Biol*, 7(7):e1002111. 39
- Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput dna sequencing. *Nucleic Acids Res*, 36(16):e105. 10, 11
- Dooner, H. K. and He, L. (2008). Maize genome structure variation: interplay between retrotransposon polymorphisms and genic recombination. *Plant Cell*, 20(2):249-258. 39

BIBLIOGRAPHY

- Durbin, R. M., Altshuler, D. L., and Gibbs, R. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073. 39
- Edgar, R. C. (2010). Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461. 15
- Ewing, A. D. and Kazazian, Jr, H. H. (2010). High-throughput sequencing reveals extensive variation in human-specific 11 content in individual human genomes. *Genome Res*, 20(9):1262–1270. 39
- 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., Fietze, 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., and Snyder, M. (2012). Architecture of the human regulatory network derived from encode data. *Nature*, 489(7414):91–100. 21
- Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., and Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *Science*, 312(5778):1355–1359. 42
- Guertin, M. J. and Lis, J. T. (2010). Chromatin landscape dictates hsf binding to target dna elements. *PLoS Genet*, 6(9). 3, 14, 35
- Haberer, G., Young, S., Bharti, A. K., Gundlach, H., Raymond, C., Fuks, G., Butler, E., Wing, R. A., Rounsley, S., Birren, B., Nusbaum, C., Mayer, K. F. X., and Messing, J. (2005). Structure and architecture of the maize genome. *Plant Physiol*, 139(4):1612–1624. 39

BIBLIOGRAPHY

- Huson, D. H., Mitra, S., Ruscheweyh, H.-J., Weber, N., and Schuster, S. C. (2011). Integrative analysis of environmental sequences using megan4. *Genome Res*, 21(9):1552–1560. 16
- Iskow, R. C., McCabe, M. T., Mills, R. E., Torene, S., Pittard, W. S., Neuwald, A. F., Van Meir, E. G., Vertino, P. M., and Devine, S. E. (2010). Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell*, 141(7):1253–1261. 39
- Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502. 1, 6
- Kaufmann, K., Muio, J. M., Jauregui, R., Airoldi, C. A., Smaczniak, C., Krajewski, P., and Angenent, G. C. (2009). Target genes of the mads transcription factor sepallata3: integration of developmental and hormonal pathways in the arabidopsis flower. *PLoS Biol*, 7(4):e1000090. 14, 35
- Kaufmann, K., Muio, J. M., sters, M., Farinelli, L., Krajewski, P., and Angenent, G. C. (2010a). Chromatin immunoprecipitation (chip) of plant transcription factors followed by sequencing (chip-seq) or hybridization to whole genome arrays (chip-chip). *Nat Protoc*, 5(3):457–472. 1, 2, 3, 4, 5, 6, 9, 10, 11, 12, 39, 41
- Kaufmann, K., Wellmer, F., Muio, J. M., Ferrier, T., Wuest, S. E., Kumar, V., Serrano-Mislata, A., Madueo, F., Krajewski, P., Meyerowitz, E. M., Angenent, G. C., and Riechmann, J. L. (2010b). Orchestration of floral initiation by apetala1. *Science*, 328(5974):85–89. 4, 10, 14, 35
- Kent, W. J. (2002). Blat—the blast-like alignment tool. *Genome Res*, 12(4):656–664. 7
- Kim, T. H. and Ren, B. (2006). Genome-wide analysis of protein-dna interactions. *Annu Rev Genomics Hum Genet*, 7:81–102. 1

- Kronmiller, B. A. and Wise, R. P. (2009). Computational finishing of large sequence contigs reveals interspersed nested repeats and gene islands in the *rf1*-associated region of maize. *Plant Physiol*, 151(2):483–495. 21, 39
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissole, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima,

BIBLIOGRAPHY

- S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and , I. H. G. S. C. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921. 39
- 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., Shores, 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., and Snyder, M. (2012). Chip-seq guidelines and practices of the encode and modencode consortia. *Genome Res*, 22(9):1813–1831. 9, 11
- Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25. 2, 7, 15
- Li, H. and Homer, N. (2010). A survey of sequence alignment algorithms for

BIBLIOGRAPHY

- next-generation sequencing. *Brief Bioinform*, 11(5):473–483. 8
- Li, H., Ruan, J., and Durbin, R. (2008a). Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–1858. 2, 7
- Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008b). Soap: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714. 2
- Li, R., Li, Y., Zheng, H., Luo, R., Zhu, H., Li, Q., Qian, W., Ren, Y., Tian, G., Li, J., Zhou, G., Zhu, X., Wu, H., Qin, J., Jin, X., Li, D., Cao, H., Hu, X., Blanche, H., Cann, H., Zhang, X., Li, S., Bolund, L., Kristiansen, K., Yang, H., Wang, J., and Wang, J. (2010). Building the sequence map of the human pan-genome. *Nat Biotechnol*, 28(1):57–63. 39
- Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., and Wang, J. (2009). Soap2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967. 7
- Longo, M. S., O’Neill, M. J., and O’Neill, R. J. (2011). Abundant human dna contamination identified in non-primate genome databases. *PLoS One*, 6(2):e16410. 12, 41
- Mallhis, N., Butterfield, Y. S. N., Ester, M., and Jones, S. J. M. (2009). Slider-maximum use of probability information for alignment of short sequence reads and snp detection. *Bioinformatics*, 25(1):6–13. 8
- Meyers, B. C., Tingey, S. V., and Morgante, M. (2001). Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res*, 11(10):1660–1676. 39
- Morohashi, K., Casas, M. I., Ferreyra, L. F., Meja-Guerra, M. K., Pourcel, L., Yilmaz, A., Feller, A., Carvalho, B., Emiliani, J., Rodriguez, E., Pellegrinet, S., McMullen, M., Casati, P., and Grotewold, E. (2012). A genome-wide regulatory framework identifies maize pericarp color1 controlled genes. *Plant Cell*, 24(7):2745–2764. 14, 35

BIBLIOGRAPHY

- Morohashi, K., Xie, Z., and Grotewold, E. (2009). Gene-specific and genome-wide chip approaches to study plant transcriptional networks. *Methods Mol Biol*, 553:3–12. 35
- Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, M., Ogasawara, N., and Kanaya, S. (2011). Sequence-specific error profile of illumina sequencers. *Nucleic Acids Res*, 39(13):e90. 10, 11
- Pareja-Tobes, P., Manrique, M., Pareja-Tobes, E., Pareja, E., and Tobes, R. (2012). Bg7: A new approach for bacterial genome annotation designed for next generation sequencing data. *PLoS One*, 7(11):e49239. 9, 16, 17
- Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for chip-seq and rna-seq studies. *Nat Methods*, 6(11 Suppl):S22–S32. 2, 6
- Richter, D. C., Ott, F., Auch, A. F., Schmid, R., and Huson, D. H. (2008). Metasim: a sequencing simulator for genomics and metagenomics. *PLoS One*, 3(10):e3373. 15
- Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M. B. (2009). Peakseq enables systematic scoring of chip-seq experiments relative to controls. *Nat Biotechnol*, 27(1):66–75. 27
- Ruffalo, M., LaFramboise, T., and Koyutrk, M. (2011). Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, 27(20):2790–2796. 41
- Rumble, S. M., Lacroute, P., Dalca, A. V., Fiume, M., Sidow, A., and Brudno, M. (2009). Shrimp: accurate mapping of short color-space reads. *PLoS Comput Biol*, 5(5):e1000386. 2, 7, 40
- Salzberg, S. L., White, O., Peterson, J., and Eisen, J. A. (2001). Microbial genes in the human genome: lateral transfer or gene loss? *Science*, 292(5523):1903–1906. 41

BIBLIOGRAPHY

- SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y., and Bennetzen, J. L. (1998). The paleontology of intergene retrotransposons of maize. *Nat Genet*, 20(1):43-45. 21, 39
- Schbath, S., Martin, V., Zytnecki, M., Fayolle, J., Loux, V., and Gibrat, J.-F. (2012). Mapping reads on a genomic sequence: An algorithmic overview and a practical comparative analysis. *J Comput Biol*. 8, 40, 41
- Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S. M., Belter, E., Du, F., Kim, K., Abbott, R. M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S. M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wisotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M. J., McMahan, L., Van Buren, P., Vaughn, M. W., Ying, K., Yeh, C.-T., Enrich, S. J., Jia, Y., Kalyanaraman, A., Hsia, A.-P., Barbazuk, W. B., Baucom, R. S., Brutnell, T. P., Carpita, N. C., Chaparro, C., Chia, J.-M., Deragon, J.-M., Estill, J. C., Fu, Y., Jeddloh, J. A., Han, Y., Lee, H., Li, P., Lisch, D. R., Liu, S., Liu, Z., Nagel, D. H., McCann, M. C., SanMiguel, P., Myers, A. M., Nettleton, D., Nguyen, J., Penning, B. W., Ponnala, L., Schneider, K. L., Schwartz, D. C., Sharma, A., Soderlund, C., Springer, N. M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T. K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J. L., Dawe, R. K., Jiang, J., Jiang, N., Presting, G. G., Wessler, S. R., Aluru, S., Martienssen, R. A., Clifton, S. W., McCombie, W. R., Wing, R. A., and Wilson, R. K.

- (2009). The b73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956):1112–1115. 21, 39
- Trapnell, C. and Salzberg, S. L. (2009). How to map billions of short reads onto genomes. *Nat Biotechnol*, 27(5):455–457. 2, 6, 7, 8
- Wang, J., Huda, A., Lunnyak, V. V., and Jordan, I. K. (2010). A gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags. *Bioinformatics*, 26(20):2501–2508. 39
- Wei, G.-H., Badis, G., Berger, M. F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A. R., Yan, J., Talukder, S., Turunen, M., Taipale, M., Stunnenberg, H. G., Ukkonen, E., Hughes, T. R., Bulyk, M. L., and Taipale, J. (2010). Genome-wide analysis of ets-family dna-binding in vitro and in vivo. *EMBO J*, 29(13):2147–2160. 30
- Xing, J., Zhang, Y., Han, K., Salem, A. H., Sen, S. K., Huff, C. D., Zhou, Q., Kirkness, E. F., Levy, S., Batzer, M. A., and Jorde, L. B. (2009). Mobile elements create structural variation: analysis of a complete human genome. *Genome Res*, 19(9):1516–1526. 39
- Zhang, Z. D., Rozowsky, J., Snyder, M., Chang, J., and Gerstein, M. (2008). Modeling chip sequencing in silico with applications. *PLoS Comput Biol*, 4(8):e1000158. 27