University of Louisville

ThinkIR: The University of Louisville's Institutional Repository

---

Electronic Theses and Dissertations

5-2018

Region based gene expression via reanalysis of publicly available microarray data sets.

Ernur Saka
University of Louisville

Follow this and additional works at: <https://ir.library.louisville.edu/etd>

Part of the [Bioinformatics Commons](#), [Computational Biology Commons](#), and the [Other Computer Sciences Commons](#)

Recommended Citation

Saka, Ernur, "Region based gene expression via reanalysis of publicly available microarray data sets." (2018). *Electronic Theses and Dissertations*. Paper 2902.
<https://doi.org/10.18297/etd/2902>

This Doctoral Dissertation is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. This title appears here courtesy of the author, who has retained all other copyrights. For more information, please contact thinkir@louisville.edu.

REGION BASED GENE EXPRESSION VIA REANALYSIS OF PUBLICLY
AVAILABLE MICROARRAY DATA SETS

By

Ernur Saka

B.S. (CEng), University of Dokuz Eylul, Turkey, 2008

M.S., University of Louisville, USA, 2011

A Dissertation Submitted To the
J. B. Speed School of Engineering
in Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy in
Computer Science and Engineering

Department of Computer Engineering and Computer Science
University of Louisville
Louisville, Kentucky

May 2018

Copyright 2018 by Ernur Saka

All rights reserved

REGION BASED GENE EXPRESSION VIA REANALYSIS OF PUBLICLY
AVAILABLE MICROARRAY DATA SETS

By

Ernur Saka

B.S. (CEng), University of Dokuz Eylul, Turkey, 2008

M.S., University of Louisville, USA, 2011

A Dissertation Approved On

April 20, 2018

by the following Committee

Dissertation Director
Dr. Eric C. Rouchka

Dr. Olfa Nasraoui

Dr. Dar-Jen Chang

Dr. Juw Won Park

Dr. Jeffrey C. Petruska

DEDICATION

Dedicated to my parents
Nuran Saka and Erol Saka
and to my sister
Esin Saka

ACKNOWLEDGEMENTS

First, I would like to thank my advisor Dr. Rouchka for his guidance, patience and support throughout my PhD. Without the guidance and the support that he has provided, this experience would not have been the same.

I would also like to thank the rest of my dissertation committee members, Dr. Chang, Dr. Park, and Dr. Petruska for giving me the opportunity to learn from them. In particular, I would like to thank Dr. Nasraoui for her constant support and encouragement. I appreciate all my professors who have helped me progress. I also would like to thank Dr. Elmaghraby, Chairman of the Department of Computer Engineering and Computer Science for his support.

My experience at the bioinformatics lab would not have been the same without our discussions, collaboration and seminars. Thanks to my former and current lab members Abdallah Eteleeb, Dazhuo Li, Fahim Mohammad, Mohammed Sayed, Mohamed Chaabane, Aanchal Malhotra, and Aryan Neupane. I would love to thank Dr. Harrison for his contributions to biological analysis and interpretation of our research outcomes. I also would like to thank Dr. Chariker for her helpful discussions and pleasant conversations we had.

Through the years, several friends have provided support. I would like to thank James Walter Moore and Caroline Fortin for always being on my side, their encouragements and listening ears.

Most of all, I am so grateful to my family, my mother Nuran, my father Erol, my sister Esin and my brother in love Levent for supporting me in every possible way unconditionally.

ABSTRACT

REGION BASED GENE EXPRESSION VIA REANALYSIS OF PUBLICLY AVAILABLE MICROARRAY DATA SETS

Ernur Saka

April 20, 2018

A DNA microarray is a high-throughput technology used to identify relative gene expression. One of the most widely used platforms is the Affymetrix® GeneChip® technology which detects gene expression levels based on probe sets composed of a set of twenty-five nucleotide probes designed to hybridize with specific gene targets.

Given a particular Affymetrix® GeneChip® platform, the design of the probes is fixed. However, the method of analysis is dynamic in nature due to the ability to annotate and group probes into uniquely defined groupings. This is particularly important since publicly available repositories of microarray datasets, such as ArrayExpress and NCBI's Gene Expression Omnibus (GEO) have made millions of samples readily available to be reanalyzed computationally without the need for new biological experiments. One way in which the analysis can dynamically change is by correcting the mapping between probe sets and targets by creating custom Chip Description Files (CDFs) to arrange which probes belong to which probe set based on the latest genomic information or specific annotations of interest.

Since default probe sets in Affymetrix® GeneChip® platforms are specific for a gene, transcript or exon, the analyses are then limited to profile differential expression at the gene, transcript or individual exon level. However, it has been revealed that untranslated regions (UTRs) of mRNA have important impacts on the regulation of proteins.

We therefore developed a new probe mapping protocol that addresses three issues of Affymetrix® GeneChip® data analyses: removing nonspecific probes, updating probe target mapping based on the latest genome information and grouping the probes into region (UTR, individual exon), gene and transcript level targets of interest to support a better understanding of the effect of UTRs and individual exons on gene expression levels. Furthermore, we developed an R package, *affyCustomCdf*, for users to dynamically create custom CDFs. The *affyCustomCdf* tool takes annotations in a General/Gene Transfer Format File (GTF), aligns probes to gene annotations via Nested Containment List (NCList) indexing and generates a custom Chip Description File (CDF) to regroup probes into probe sets based on a region (UTR and individual exon), transcript or gene level.

Our results indicate that removing probes that no longer align to the genome without mismatches or align to multiple locations can help to reduce false-positive differential expression, as can removal of probes in regions overlapping multiple genes. Moreover, our method based on regions can detect changes that would have been missed by analysis based on gene and transcript. It also allows for a better understanding of 3' UTR dynamics through the reanalysis of publicly available data.

TABLE OF CONTENTS

DEDICATION	iii
ACKNOWLEDGEMENTS	iv
ABSTRACT	vi
LIST OF TABLES	xiii
LIST OF FIGURES	xv
1 INTRODUCTION	1
1.1 Motivation	2
1.2 Organization of the Document	8
2 OVERVIEW OF MOLECULAR BIOLOGY	10
2.1 DNA	11
2.2 RNA	15
2.3 mRNA Structure	19
2.3.1 Untranslated Regions (UTR)	19
2.3.2 Poly(A) Tail	20
2.4 Protein	20
2.5 Chromosome	21
2.6 Gene	22
2.7 Gene Regions	22
2.3.1 Intron.....	22

2.3.2 Exon	23
2.3.3 Open Reading Frame (ORF)	23
2.3.4 Coding sequence (CDS)	23
2.8 Gene Expression	24
2.9 Central Dogma of Molecular Biology	25
2.9.1 General Transfers	26
2.9.2 Special Transfers	29
2.10 Genetic Code	30
2.11 Alternative Splicing	31
2.12 Transcript	33
2.13 Noncoding DNA	33
2.13.1 Promoter	34
2.13.2 Enhancer	34
2.13.3 Silencers	34
3 MICROARRAYS	35
3.1 DNA Microarray	36
3.1.1 Hybridization	37
3.1.2 DNA Microarray Experiment	37
3.1.3 Experimental Variations in Microarray Platforms	41
3.1.4 Types of the DNA Microarrays Based on Technology	41

3.2 Affymetrix® GeneChips®	43
3.2.1 Array Design	45
3.2.1 Attaching Probes	48
3.3 Data Files.....	49
3.3.1 Affymetrix GeneChip® Data Files	49
3.3.2 FASTA File Format	53
3.3.3 General/Gene Transfer Format	54
4 CUSTOM CHIP DESCRIPTION FILES (CDF)	56
4.1 Limitations of Microarrays.....	57
4.2 Custom Designed CDF Files.....	60
5 COORDINATE MAPPING BETWEEN PROBES AND ANNOTATIONS	75
5.1 Bowtie	76
5.1.1 Burrows-Wheeler Transform	77
5.1.2 Inexact Matching	78
5.1.3 Double Indexing.....	79
5.2 Interval	80
5.3 Interval Tree	81
5.3.1 Interval Tree Operations	83
5.4 Nested Containment Lists	84
5.4.1 NCList Data Structure.....	85

5.4.2 Query of overlapping intervals	86
6 METHODS	88
6.1 Steps for Generating Custom CDFs	88
6.1.1 Mapping of Perfect Match Probes to a Genome	89
6.1.2 Annotation of Perfect Match Probes via NCList	91
6.1.3 Probe Set Naming	97
7 affyCustomCdf: AN R PACKAGE FOR CREATING CUSTOM CDF	98
7.1 affyCustomCdf Tool	98
7.1.1 Custom CDF Types.....	98
7.1.2 Interpreting Results	101
7.1.3 Running Time	102
8 EXPERIMENTAL RESULTS	104
8.1 Probes Mapping to Genome	104
8.1.1 HG-133 Plus 2.....	105
8.1.2 Rat 230 2.0.....	105
8.1.3 Mouse 430 2.0.....	106
8.2 Probe Annotations and Probe Sets	106
8.2.1 Detailed Results for HG-133 Plus 2.....	107
8.3 Analysis Results	113
8.4 Conclusions	128

9 CONCLUSIONS AND FUTURE DIRECTIONS	131
9.1 Summary of Contributions	132
9.2 Post Dissertation Work.....	133
9.2.1 Dealing with Dynamic Biological Initializers	133
9.2.2 Conservation of Gene Annotation	133
9.2.3 Disease Drug Network Establishment Based on Genomic Expression Profiles	135
9.2.4 Untranslated Regions Gene Expression Profiling.....	136
REFERENCES	138
APPENDIX	144
CURRICULUM VITA	151

LIST OF TABLES

1	Release dates of databases used by NetAffx v36 annotations and current database versions	4
2	Top Affymetrix® in situ oligonucleotide arrays found in GEO	4
3	Alternative CDFs for the top Affymetrix® in situ oligonucleotide arrays found in GEO	4
4	Genetic code	31
5	Transcripts of the Gene CTLA4	33
6	The release date and the number of table human sequences for UniGene, dbEST, Ensembl and RefSeq at the time of Affymetrix® HG-U133 design and the current versions	58
7	The alignment of PM probes taken from HG 133 plus-2 array PM to RefSeq database together with their corresponding gene symbol and the designated gene taken from gin file. The alignments were achieved via Absolute Gene ID Conversion Tool	60
8	Summary of available custom CDFs done by several researchers	61
9	Probe annotation outputs at the end of NCList indexing search.....	93

10	Custom CDF naming examples	97
11	Running times	102
12	Number of mapped probes for custom CDF Construction	104
13	Summary of probes used for custom gene and transcript based Custom CDFs	107
14	Summary of probes used for region based custom CDFs.....	107
15	DS1 versus DS2U top ten DEGs	115
16	DS4 versus DS2U top ten DEGs	116
17	DS1 versus DS2U top ten region based DEGs	118
18	DS4 versus DS2U top ten region based DEGs	119
19	D1 versus DS2U top ten upregulated genes Brain Array Ensembl V19	120
20	Day 7 versus naive top ten DEGs	122
21	Day 14 versus naive top ten DEGs	122
22	7 versus naive top ten region based DEGs.....	125
23	14 versus naive top ten region based DEGs.....	125
24	The number of probes per region and transcript after the annotation step	134

LIST OF FIGURES

1	The relationship among the cell, the nucleus, the chromosome, the gene and DNA ..	11
2	The basic unit of polynucleotide chain is the nucleotide	12
3	Types of nitrogenous bases	13
4	Adjacent nucleotides connect to form a DNA polynucleotide	13
5	Semantic representation and sequence of a single strand DNA	14
6	Complementary base pairs	14
7	Semantic representation and sequence of complementary strands	14
8	DNA double helix	15
9	DNA to pre-mRNA	16
10	DNA to mRNA	17
11	tRNA structure	18
12	Mature mRNA	19
13	Simplified diagram of protein production from DNA	21
14	Possible open reading frames from a DNA sequence	24

15	Central Dogma of Molecular Biology	26
16	DNA replication.....	27
17	Transcription	28
18	Translation	29
19	Alternative splicing types	32
20	Microarray features and probes	36
21	One-color Affymetrix® GeneChips® microarray experiment diagram.....	38
22	A typical two-color microarray experiment diagram.....	39
23	Affymetrix® GeneChips® microarray design.....	44
24	Affymetrix® GeneChips® probe selection and perfect match (PM), mismatch (MM) probe sets	44
25	Affymetrix® GeneChips® Human Genome U133 multiple probe selection regions .	46
26	GeneChips® Human Genome U133 arrays transcript, consensus sequence, and probes relationship	46
27	GeneChips® Human Genome U133 Plus-2 probe set marking	48
28	Affymetrix GeneChip microarray probe attaching via masked photolithography	49
29	Part of the ASCII EXP file of chicken genome array supplied by Affymetrix® in the Sample Data for GeneChip® Arrays	50
30	Part of the ASCII CEL file of chicken genome array supplied by Affymetrix® in the Sample Data for GeneChip® Arrays	51
31	Part of the ASCII HG-U133_Plus_2.cdf file	52

32	Part of the ASCII HG-U133_Plus_2.gin file	52
33	Part of the ASCII HG-U133_Plus_2.PSI file	53
34	A part of the Affymetrix® GeneChip® HG-133 Plus 2 FASTA file	53
35	A part of the Ensembl gene annotation GTF file for human build 38	55
36	A Burrows-Wheeler transform	77
37	Exact matching versus inexact alignment.....	80
38	An interval tree example. (a) A set of intervals. (b) The interval tree constructed from the set of intervals	82
39	Interval tree search pseudo code	83
40	Difference between start coordinate sorted database and a NCList data structure. Intervals shown in bold are overlapping intervals with the query interval	85
41	Schematic representation of NCList data structure. Here in the interval array L, element L[i] has non-empty sublist, which records is indexed in sublist header array H by L[i]. SUBLIST. This sublist starts at START and has length LENGTH. L[i] itself belongs to the top level list S0, which has sublist header H[0]	86
42	NCList overlap query pseudo code.....	87
43	Flow chart for region-based probe annotation framework	90
44	Features of the FO538757.2 gene in GTF	92
45	Creating probe sets for different types of custom CDF based on probe mapping to genomic structures	96
46	affyCustomCdf parameter selection	103

47	The distributions of probe numbers per probe set for the gene region based CDF ...	108
48	Original CDF versus custom CDF	109
49	The distributions of probe numbers per probe set for the gene based CDF	111
50	Original CDF versus the brain array CDF	111
51	The distributions of probe numbers per probe set for the transcript based CDF	112
52	Typical method versus our method for DS1 versus DS2U	114
53	Typical method versus our method for DS4 versus DS2U	115
54	Distributions of probe set regions in the DS1 versus DS2U DEGs	116
55	Distributions of probe set regions in the DS4 versus DS2U DEGs	117
56	Distribution of common and different DEGs for our gene and gene region based methods in the DS1 versus DS2U case	117
57	Distribution of common and different DEGs for our gene and gene region based methods in the DS4 versus DS2U case	118
58	DEGs detected by the Brain array method and our gene based method for DS1 versus DS2U	119
59	Typical method versus our method for day 7 versus naive	121
60	Typical method versus our method for day 14 versus naive	121
61	Distributions of probe set regions in the 7 versus naive DEGs	123
62	Distributions of probe set regions in the 14 versus naive DEGs	123
63	Distribution of common and different DEGs for our gene and gene region based methods in the 7 versus naive case	124

64	Distribution of common and different DEGs for our gene and gene region based methods in the 7 versus naïve case	124
65	Number of common and different differentially expressed genes using region, gene and brain array custom CDFs. a Day 7 versus naïve. b Day 14 versus naïve.....	127
66	Probe set expression levels within the gene, exon, and 3' UTR regions for 12 different genes	130
67	Genome browser view of the gene Camk4 with probes	134
68	Genome browser view of the gene Otud7a with probes	135
69	First ten lines of the Rat 230 2.0 microarray probe mapping file	145
70	Bowtie command line for probe mapping	145
71	Mapping result of Bowtie for Rat 230 2. microarray.....	142
72	Example R code for clearing Bowtie mapping results.....	146
73	File naming for custom CDFs.....	148
74	Calling createAffyCustomCdf function	150

CHAPTER 1

INTRODUCTION

Photo 51, the first X-ray diffraction image of DNA produced by Rosalind Franklin in 1952, triggered the discovery of the DNA helix structure by James D. Watson and Francis H. C. Crick in 1953. Subsequently, the first DNA polymerase was identified by Arthur Kornberg in 1957. After the discovery of the DNA structure and the DNA replication mechanism, researchers started working on gene expression analysis and protein synthesis [1], leading to the current day fields of molecular biology and bioinformatics.

Gene expression analysis studies the entire process of a particular worker molecule production from the coding information contained in DNA. Technologies developed for gene expression studies have had a huge impact on the field. Southern blotting developed by Edward M. Southern and Sanger sequencing developed by Frederick Sanger enabled researchers to locate particular gene or DNA samples in the genome. Northern blotting developed by James Alwine and George Stark allowed detection of specific mRNA in a geometric sample. Exponential amplification of DNA segments became possible via the polymerase chain reaction (PCR) technique developed by Kary Mullis. The development of microarrays allowed researchers to investigate thousands of gene products in parallel.

One of the most widely used microarray platforms is the Affymetrix® GeneChip® family of arrays. It detects gene expression levels based on probe sets composed of a set of individual 25 base probes designed to hybridize with the specific gene targets. This research mainly focuses on the creating a framework for computationally reanalysis of publicly available Affymetrix® GeneChip® data based on region (UTR and individual exon), gene and transcript through remapping probes to gene targets with created custom Chip Description Files (CDFs). The custom CDF creation involves removing nonspecific probes, updating probe target mapping based on the latest genome information, grouping probes into region, gene or transcript level targets and saving updated probe-target groups into a CDF. The unspecified probes were identified based on the mapping of probe sequences to the genome of interest. Probe target mapping was accomplished by aligning specific probes to genomic intervals obtained from general/gene transfer format file via Nested Containment List (NCList) indexing. Our custom CDFs provide a way to investigate changes appeared in the untranslated regions and exons of mRNA which data has not been performed on a large scale with microarray datasets in addition to gene and transcript. We also supply flexibility for creating gene of interest CDFs by allowing user to supply annotations via GTF files.

1.1 Motivation

Our research has been motivated by the following goals:

- 1) Improving the accuracy of microarray by using up-to-date genomes and annotations.

Given a particular Affymetrix® GeneChip® platform, the design of the probes is fixed based on earlier genome assemblies and annotation available at that time. Since the

design of the first Affymetrix® GeneChip®, rapid progress has been made in genome sequencing resulting in more accurate databases of annotated protein coding and non-coding structural genes. The significant differences between old and new genome assemblies and annotations make it necessary to update probe-gene targeting according to current knowledge to get more accurate interpretations from experimental results. Affymetrix® attempts to provide compatibility between genomic changes by updating links between probe sets and their corresponding genes/transcripts via NetAffx™[2]. Table 1 shows release dates of source databases used by Affymetrix® version 36 for both the incorporated version and the most recently available version. In all cases, there is at least three-year difference between the incorporated and most recent release dates which can lead to inconsistent interpretation.

In addition, updating links between probe sets and their corresponding genes/transcripts does not provide a solution for problems caused by individual probes such as single nucleotide polymorphisms (SNPs) [3, 4] probes that target genes other than the designated gene of a probe set, and probes that no longer align to a genomic location due to refinements in genome assemblies. For example, in the Affymetrix® GeneChip® HG-133 Plus 2 array, a total of 40,680 probes out of 603,158 (excluding quality control probes) do not have a perfect match to the most recent human genome assembly (hg38).

Although the inherent effects of using dated probe gene mapping designs to analyze microarray data sets might seem obvious, the overwhelming majority of experimental results have only been analyzed using the original CDFs designed by Affymetrix®. For example, as of May 2016, GEO has 120,920 samples which were analyzed via the original Affymetrix® CDFs for the HG-U133 Plus 2 array (Table 2). On

the other hand, only 6,403 samples were analyzed using custom CDFs, mostly produced by brainarray (Table 3). Given that fewer than 5% of all samples in GEO have been analyzed by alternative CDFs, an opportunity exists to reanalyze existing datasets according to updated transcript knowledge or functional regions of interest.

TABLE 1: Release dates of databases used by NetAffx v36 annotations and current database versions

GEO Platform	Organism			
GPL570	<i>Homo sapiens</i>			
GPL1261	<i>Mus musculus</i>			
GPL1355	<i>Rattus norvegicus</i>			
GPL198	<i>Arabidopsis thaliana</i>			
Incorporation Data of Databases Common to All Four GEO Platforms				
	Ensembl	RefSeq	GenBank	Entrez Gene
NetAffx	Nov-15	Nov-15	Nov-15	Nov-15
Current	Apr-18	Mar-18	Feb-18	Apr-18

TABLE 2: Top Affymetrix® in situ oligonucleotide arrays found in GEO

GEO Platform	Title	# of Probes	# of Probe Sets	# of Samples
GPL570	Human Genome U133 Plus 2.0 Array	604,258	54,675	120,920
GPL1261	Mouse Genome 430 2.0 Array	496,468	451,01	480,87
GPL1355	Rat Genome 230 2.0 Array	342,410	310,99	189,12
GPL198	Arabidopsis ATH1 Genome Array	251,078	228,10	126,24

TABLE 3: Alternative CDFs for the top Affymetrix® in situ oligonucleotide arrays found in GEO

GEO Platform	# of Alternative CDFs	# and percent of Samples Using Alternative CDFs
GPL570	54	6403 (5.0%)
GPL1261	36	1984 (4.0%)
GPL1355	12	460 (2.4%)
GPL198	9	642 (4.8%)

- 2) Understanding the effect of untranslated regions on gene expression levels furthermore on diseases.

In molecular biology, even though the scaffolds of the proteins are defined by genomic DNA during transcription, they are actually synthesized during translation based on the intermediary template messenger mRNA using the genetic code. But regulation of genes, which functions in the selection of produced proteins, can occur at the transcriptional or translational level [5]. Transcription level regulation occurs through complexes that form at transcription factor binding sites which typically found within 5' UTRs, the upstream region of the transcription start site (TSS). On the other hand, 3' UTRs, which are downstream of the coding region, serve as a binding site for many translational control mechanisms.

While microarrays have been successfully utilized for understanding differential expression at the gene or probe set level, less attention has been given to the potential analysis at the individual exon, alternative transcript, and untranslated region (UTR) level. Although the selection bias of probes on the 3' ends of genes for earlier iterations of Affymetrix® GeneChip® designs presents limitations on the completeness of transcript information, more recent designs allow for a more complete coverage of exons and exon junctions. However, information concerning individual exons can still be extracted from earlier GeneChip® designs, particularly in the 3' UTR regions.

- **3' Untranslated Regions (UTRs) association with gene regulation**

3' UTRs contains translational control mechanisms that include miRNA binding sites, AU rich elements (AREs), cytoplasmic polyadenylation element (CPE) binding sites,

localization binding elements, zipcode binding proteins and G-quadruplex sites. Through these mechanisms, 3' UTRs play important role in development [6-10] embryonic axis formation, neurogenesis, erythropoiesis, localization in the nervous system [11-16] and cancer [17-19]. For example, cytoplasmic polyadenylation [20], which is a modification of mRNA transcripts in the form of polyadenylated (poly(A)) tails, plays an important role in transcription level regulation, indicating that alternative splicing in the 3' UTR might be associated with the changes in the coding regions thus structural changes within the 3' UTR will affect the expression level and localization of a gene. Over 40% of genes have been shown to generate multiple mRNAs with variable 3' UTR lengths [21]. These 3' UTRs harbor binding sites for molecules including microRNAs (miRNAs) and RNA-binding proteins. Thus, mRNA isoforms with lengthened 3' UTRs have increased numbers of sites for these cis-interacting factors. The diversity of 3' UTRs is predominantly regulated by alternative polyadenylation (APA), which employs alternative mRNA cleavage sites that lie progressively distal to the stop codon. APA-driven mRNA diversity is required for normal physiology, and misregulation of this process is associated with diverse disease state [22]. It is known that the length of the 3' UTR (short and long 3' UTRs) plays an important role in localization of brain-derived neurotrophic factor (*BDNF*) transcripts. Short 3' UTR mRNAs are often localized in somata, while long 3' UTR mRNAs localized in dendrites [23]. Lengthening of 3' UTRs has an effect during development. A study has shown that in the mouse brain during embryonic and postnatal development, mRNAs have longer 3' UTRs than other tissues [7]. Shortening of 3' UTRs also plays an important role. It has been shown that tumors with shorter 3' UTRs are more aggressive in nature, and 3' UTRs expression signatures are being used as strong predictors of survival [24]. The

mutations within the 3' UTR affect the termination codon, polyadenylation signal and change the secondary structure of the 3' UTR. For example, aniridia disease resulting partial in or complete loss of the iris is associated with the mutation in the stop codon (TAA → TTA) located within the 3' UTR of the *PAX6* gene [25].

- **5' Untranslated Regions (UTRs) association with gene regulation**

The 5' UTR is a regulatory region located at the 5' end of the mRNA [26]. Some of the 5' UTR regulatory elements which cause abnormal cell function are 5' UTR length, 5' cap structure, secondary structure, mutation in the 5' UTR and uORFs (upstream open reading frames). For example it was shown that the *BRCA1* gene, which acts as a tumor suppressor, is downregulated by a G to C mutation within the 5' UTR of the gene itself [25].

3) Taking advantage of available data in public repositories.

Since microarrays were introduced in 1995, different platforms have been developed and used to perform lab experiments. Over two million different samples have been produced and submitted to publicly available repositories. One of the most known and used repositories in international public high-throughput functional genomics is the Gene Expression Omnibus (GEO). As of April 2018, GEO has 2,455,165 samples, 97,014 series and 4,348 data sets. Each sample contains original experiment results under the specific conditions supplied by submitter. A series is a group of related samples with the summary of the experiment, conclusions and analysis. A data set is composed of curated records produced by GEO staff by reassembling biologically and statistically comparable GEO samples.

In microarray technology, even though the design of the probes is fixed, the methods with which the resulting experiments can be analyzed are dynamic in nature due to the ability to annotate and arrange probes into uniquely defined groupings. This is particularly important since GEO contains 29,930 samples produced by Affymetrix® GeneChips® platform as of April 2018 that can be reanalyzed computationally based on current knowledge without the need for new biological experiments. As a case in point, each of the four most commonly used species have samples that have been analyzed using the original CDFs (Table 2).

1.2 Organization of the Document

This dissertation is organized in 9 chapters. Chapter 2 is a brief overview of molecular biology. It begins with an introduction followed by the main genetic molecules DNA, RNA and protein. Next it explains gene, gene regions and the gene expression mechanism. It continues with the central dogma, genetic information transfer and genetic code. It ends with alternative splicing and noncoding DNA regions. Chapter 3 explains microarrays which mainly focuses on the Affymetrix® GeneChip® platform. It starts with a history of gene expression tools before microarrays and continues with the basis of microarrays; hybridization, design of DNA microarray lab experiments and type of DNA microarray technologies. Next it explains the Affymetrix® GeneChip® technology and ends with the data file types. Chapter 4 is the literature review of custom CDFs. It starts with a motivation for the used custom CDFs and summaries the custom CDF design steps of previously published works. Chapter 5 explains the alignment tool Bowtie, the interval tree data structure and nested containment list indexing. Chapter 6 explains the protocol used to create our custom CDFs. It starts with the removal of unspecific probes, continues

with annotation of probes via NCList indexing and ends with probe set naming inside the custom CDF file. Chapter 7 explains the developed R package `affyCustomCdf` and a variety of files that can be created via the `affyCustomCdf` tool. Chapter 8 gives the statistical and analysis results of produced human, rat and mouse custom CDFs. It also presents the comparisons between our results, previous work and comparisons between different types of our custom CDFs via reanalyzing publicly available microarray datasets. Chapter 9 discusses the contributions and post dissertation work. Appendix A provides steps for using the developed R package `affyCustomCdf`.

CHAPTER 2

OVERVIEW OF MOLECULAR BIOLOGY

Molecular biology is the field of study that involves studying cell structure and function down to the level of the individual molecules within those cells that contain the programming for life functions [27]. Cells are the smallest living things and the basic unit of any living organisms. They have all the properties of life such as reproduction, response to environment signals, a need for energy, and release of waste products. All cells are built out of similar materials (including an organism's DNA) and function in similar ways. Prokaryotic cells, such as bacteria and some single celled organisms, have simple organization. They do not have a true membrane-bound nucleus and organelles. Eukaryotic cells, such as plants, animals, and fungi are structurally complex. They contain a membrane-bound nucleus and organelles.

The nucleus is a small spherical, dense body in a eukaryotic cell. It is called the control center of the cell since it controls many of the activities of the cell including cell reproduction. Contained within the nucleus are chromosomes which are microscopic, threadlike strands composed of DNA. Regions of the DNA are gene coding segments used by a cell to create cellular workers like proteins that control the function of a cell. The proteins are coded by the sequence of DNA which is written in the chemical letters A, T, C, and G. When proteins are needed, the information contained in the DNA is transcribed

into RNA. The RNA is first processed and then transported out of the nucleus. Outside the nucleus, the proteins are built based upon the code in the RNA. Fig. 1 shows the relationship between nucleosome, chromosome, genes, and DNA.

Figure 1: The relationship among the nucleus, the chromosome, the gene and DNA [28]

2.1 DNA

Deoxyribonucleic acid (DNA) is a long-term storage device of organisms to store genetic information. It enables the transmission of genetic material from one generation to the next by passing copies of DNA to the offspring. Stored information is read by working

cells to build molecules, such as protein and RNA. These molecules are used to control how an organism looks, behaves and reproduces.

DNA may be single or double stranded. A single stranded DNA molecule, called a polynucleotide, is a chain of small molecule nucleotides. Each nucleotide is made with three separate parts: a phosphate, sugar, and nitrogenous base (Fig. 2).

Figure 2: The basic unit of polynucleotide chain is the nucleotide

I. Phosphate component: The phosphate group is a phosphorus atom surrounded by four oxygen atoms. When nucleotides are joined together to form a polynucleotide, a phosphate group is attached to the sugar molecule of adjacent nucleotide to form the sugar phosphate backbone.

II. Sugar component: There are two different kinds of sugars found in a nucleotide: deoxyribose and ribose. In DNA the sugar component of the nucleotide is deoxyribose.

III. Nitrogenous base: There are five different canonical nucleotide bases. These bases are Adenine (A), Guanine (G), Cytosine (C), Thymine (T) and Uracil (U) (Fig. 3). DNA is composed of the four bases adenine, guanine, cytosine and thymine. These five bases can be put into two categories: purine and pyrimidine.

Adenine and guanine are purines and have similar structure. Cytosine, thymine and uracil are pyrimidines and have a smaller structure than the purines.

In joining nucleotides together, the sugar part of one nucleotide connects up to the phosphate part of the next nucleotide to produce a polynucleotide (Fig. 4).

Figure 3: Types of nitrogenous bases

Figure 4: Adjacent nucleotides connect to form a DNA polynucleotide

A polynucleotide can be any length and have any order. The end of the polynucleotide is marked either 5' or 3' representing the location of the hydrogen bond.

DNA is usually written with 5' left and 3' right. Fig. 5 shows the semantic representation and sequence of a single strand DNA.

Figure 5: Semantic representation and sequence of a single strand DNA

DNA is typically a double-stranded molecule, consisting of two complementary strands running in opposite directions. One chain runs 5'-3' and the other runs 3'- 5'. The two strands are connected to each other by hydrogen bonds pairing each base on one strand to a specific partner on the other strand. These complementary base pairs are adenine (A) - thymine (T), and guanine (G) - cytosine (C) (Fig. 6). Referring to the complementary base pairing mechanism, the second strand is known as the reverse complement of the first strand (Fig. 7). Complementary base pairing helps achieve direct synthesis of a complementary strand by using one strand of DNA as a template to copy the second strand.

Figure 6: Complementary base pairs

Figure 7: Semantic representation and sequence of complementary strands

Two complementary DNA strands are twisted to make a double helix structure (first discovered by Watson and Crick[29]). In the helix structure, the four nucleotides make up the stairs and strands run in opposite directions.

Figure 8: DNA double helix [30]

2.2 RNA

Ribonucleic acid (RNA) is very similar to DNA; however a few important structural details are different. RNA is usually a single stranded molecule, while DNA is usually double stranded. RNA nucleotides contain ribose as the sugar component while

DNA nucleotides contain deoxyribose. Adenine, guanine, and cytosine are common bases for both RNA and DNA. But RNA uses the nucleotide uracil, instead of thymine.

RNA plays a key role in the Central Dogma of Molecular Biology [31] which is a pathway from DNA to proteins. Three kinds of RNA molecules perform different but cooperative functions in protein synthesis. These are:

I. Messenger RNA (mRNA): mRNA is a single strand RNA molecule used to transfer genetic information from DNA to ribosome via transcription. Because the information in DNA cannot be decoded directly into proteins, DNA is first encoded into mRNA during transcription.

Transcription of DNA to mRNA is a two-step process. First, pre-mRNA is synthesized from one strand of a DNA using a complementary mechanism (Fig.9).

Figure 9: DNA to pre-mRNA

Pre-mRNA contains both non-coding regions (introns) and coding regions (exons). Only exons are used to build proteins. Therefore, in the second step the introns are removed and the exons are spliced together to form mRNA (Fig.10). Later the formed mRNA is carried from the nucleus to the cytoplasm to be translated into a protein sequence.

The amount of mRNA produced from DNA gives a measurement of the activity of individual genes in a cell, since only active genes are translated to mRNA.

II. Ribosomal RNA (rRNA): Ribosomal RNA (rRNA) is a non-coding ribonucleic acid that is an essential and functional component of ribosomes. Ribosomes are small particles located in the cytoplasm (jelly like material that fills the cell).

As non-coding RNA, rRNA itself is not translated into a protein, but provides a mechanism for translating mRNA into protein by interacting with the transfer RNAs during translation.

Figure 10: DNA to mRNA

III. Transfer RNA (tRNA): Transfer RNA (tRNA) is a specialized RNA that is produced by transcription like mRNA. tRNA functions in carrying amino acids to the ribosome to

form proteins. tRNA has a unique three-dimensional structure that helps to perform their function (Fig. 11).

Figure 11: tRNA structure [32]

Each tRNA molecule has an anticodon (three nucleotide sequence) and amino acid attachment site. The anticodon is unique for each amino acid which means that each tRNA binds a specific amino acid. During the translation of mRNA sequence to an amino acid chain (protein), the anticodon temporally pairs with a complementary consecutive triplet, the codon, in mRNA. At the same time the amino acid binds to tRNA by the help of enzymes. Subsequently the bound amino acid is transferred to the ribosome, where proteins are assembled according to the information carried by mRNA, and attached to the growing amino acid chain.

2.3 mRNA Structure

A processed mRNA consists of different regions as shown in Fig. 12.

Figure 12: Mature mRNA

2.3.1 Untranslated Regions (UTR)

UTRs are non-coding regions located on each side of a coding sequence on an mRNA.

- **5 Prime UTR (5' UTR):** The section of mRNA from the 5' end of the mRNA to just before the first codon used in translation. It functions as a transcriptional regulator [33].
- **3 Prime UTR (3' UTR):** The region on the mRNA from the 3' end of the mRNA. 3' UTRs affect the produced gene expression level by the mRNA. Prior to translation microRNAs can bind to the 3' UTR and reduce the gene expression level of mRNAs by marking it for degradation [34]. Silencer regions in the 3' UTR inhibit gene expression. In addition, proteins can bind to the adenylate-uridylate-rich elements (AU-rich elements, AREs) located in the 3' UTR region and effect stability and degradation rate of transcripts. Moreover, poly(A) signals contained in the 3' UTR controls the addition of a poly(A) tail, which can be several hundred adenine bases, to the 3' end of the mRNA.

2.3.2 Poly(A) Tail

When transcription of a gene terminates, a long chain of adenine bases called poly(A) tail is added to the 3' end of a messenger RNA [20]. It protects the mRNA molecule from enzymatic degradation, makes the mRNA more stable, and allows mRNA to be exported from the nucleus into the cytoplasm where it is translated into protein.

2.4 Protein

Proteins are the biological molecules necessary for most of the activities in organisms including DNA replication, transportation of molecules and catalyzing biochemical reactions (enzymes) [35]. A protein molecule is made from long chain of amino acids. An amino acid is an organic compound which contains amine (NH_2) and carboxylic acid (COOH) functional groups, along with a side-chain or R-group what makes each amino acid different from the others.

Proteins vary in their amino acid sequence which is encoded in the nucleic acid sequence of their gene. A protein is formed by a multiple step process (Fig. 13). A gene in DNA that carries the encoding of a specific protein is first transcribed into pre-messenger RNA (mRNA) via proteins such as RNA polymerase. Later pre-messenger RNA is processed to form mature mRNA and alternatively spliced to create different protein isoforms from a gene. The resulting mRNA is translated into protein by the process called translation. During translation, mRNA is read according to the genetic code. The genetic code is the set of rules for coding amino acids from three consecutive nucleotides called codons. For example, CAG (cytosine-adenine-guanine) codes for glutamine (Gln).

Figure 13: Simplified diagram of protein production from DNA

A protein folds into a specific three-dimensional structure defined by the amino acid sequence. The three dimensional structure determines the function of the protein since proteins function via physical interaction with other molecules [36].

2.5 Chromosome

In a single human cell there are 46 chromosomes in DNA that makes 3 billion base pairs of DNA per cell. Because each base pair is around 0.34 nanometers long, each diploid cell contains about 2 meters of DNA. Moreover, an adult human body has about 50 trillion cells, which means there is 100 trillion meters of DNA per human. Since the sun is 150

billion meters away from earth, each human has enough DNA for more than 300 trips from the earth to the sun and back [37]. To fit long DNA strands into the microscopic space of a cell nucleus, DNA is packed into structures called chromosomes. The packing of DNA into a chromosome is done in several steps, starting with the double helix structure of DNA. Then, DNA is wrapped around proteins called histones. The resulting DNA protein complex is called chromatin. The fundamental packing unit of chromatin is the nucleosome. The nucleosome must be stable and tightly bound to compact DNA but at the same time must allow access to the DNA for the regulatory control to ensure correct gene expression. Eventually nucleosomes are folded and form chromosomes (Fig. 1).

2.6 Gene

A gene is a continuous subpart of a single stranded DNA molecule. One strand of DNA contains many genes. All of these genes are needed to give instructions for how to build all of the proteins for an organism. Fig. 1 shows the relationship between a gene and DNA.

2.7. Gene Regions

A gene consists of different regions and can be classified as protein coding genes or non-coding RNA genes. A nucleotide sequence in the gene may be in one of the following regions:

2.7.1 Intron

An intron is a non-coding nucleic acid sequence region. It is located inside of a gene and the corresponding RNA transcript of most of the organisms. The amount and size of introns varies according to organism, gene and location of a gene in the cell. They are removed during alternative splicing and do not have an effect on the final product of the

gene. But it does not mean that they do not serve a purpose. After alternative splicing they may produce non-coding RNA molecules [38]. Furthermore, some introns encode specific proteins by themselves.

2.7.2 Exon

An exon is a coding nucleic acid sequence part of a gene. It contains information for protein synthesis. After introns are removed, exons join together during alternative splicing and create mature mRNA.

2.7.3 Open Reading Frame (ORF)

An open reading frame is the part of the reading frame that starts with start codon ATG (Met) and ends with one of the three stop codons (TAA, TAG, TGA). A reading frame is a sequence of nucleotide triplets (codons) which contains directions for making a protein. In a single strand of nucleic acid there are three possible reading frames, each beginning from a different nucleotide in a triplet.

In double stranded DNA, an additional three possibilities come from the complementary sequence. Gene prediction starts with an open reading frame search.

2.7.4 Coding sequence (CDS)

CDS is a region of DNA that is known to be translated to a protein. It is also called the coding sequence. A CDS begins with start codon ATG next to the 5' end and terminates with one of the three stop codons (TAA, TAG, TGA) next to 3' end. The main difference of CDS from ORF is they are known to be transcribed into a protein, neither the gene nor the protein must be known but ORFs are potential protein coding regions and may also contain non-coding RNA.

5' ... agtacgctgaatagcgtattttcatttgaggacgatgataaa ... 3'	
Frame 1	agt acg ctg aat agc ttt tca ttt gag gac gat gta taa
Protein Sequence	S T L N S F S F E D D V *
Frame 2	gta cgc tga ata gcg tat ttt cat ttg agg acg atg tat
Protein Sequence	V R * I A Y F H L R T M Y
Frame 3	tac gct gaa tag cgt att ttc att tga gga cga tgt
Protein Sequence	Y A E * R I F I * G R C
The sample DNA sequence is showing 3 different reading frames in the forward direction. Open frame 1 starts with “a”, open frame 2 starts with “g”, open frame 3 starts with “t” and “*” indicates the stop codons. Each frame line followed by translated amino acid sequence. The possible longest ORF is in open frame 1.	

Figure 14: Possible open reading frames from a DNA sequence

2.8 Gene Expression

Almost every cell in an organism contains a complete set of genes but each gene in the set is not used by a specific cell due to the effect of cell type, cell development and environmental changes. This important mechanism of cells relates the importance of differential gene expression.

The separation of active and inactive genes is carried out by a process known as gene regulation. When a gene is turned on, the molecular product of this gene can be synthesized, and subsequently identified as expressed. Oppositely when a gene is turned off, the molecular product of this gene cannot be synthesized and the gene is identified as unexpressed.

Gene expression analyses reveals the function of genes, cell-cell differences, cell interactions and where, when and in which conditions a gene expressed. The expression

analyses of many genes can be determined by measuring mRNA levels with multiple techniques including in situ hybridization and microarrays.

2.9 Central Dogma of Molecular Biology

In 1958, Francis Crick used the term “Central Dogma” for the idea of one-way genetic information flow between macromolecules and he explained it as “once ‘information’ has passed into protein it cannot get out again. In more detail, the transfer of information from nucleic acid to nucleic acid or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible” [31].

In the Central Dogma there are three major classes of macromolecules: DNA, RNA (both nucleic acids) and proteins [39]. By using the three macromolecules, nine information transforms are defined and classified into three groups: general, special and unknown transfers. The general transfers are: $\text{DNA} \rightarrow \text{DNA}$, $\text{DNA} \rightarrow \text{RNA}$ and $\text{RNA} \rightarrow \text{Protein}$ transforms. They usually occur in most cells. The special transfers are: $\text{DNA} \rightarrow \text{Protein}$, $\text{RNA} \rightarrow \text{DNA}$ and $\text{RNA} \rightarrow \text{RNA}$. They only occur in a laboratory or in the case of some viruses under specific conditions. The unknown transfers are: $\text{Protein} \rightarrow \text{DNA}$, $\text{Protein} \rightarrow \text{RNA}$ and $\text{Protein} \rightarrow \text{Protein}$. Fig. 15 is taken from original work of Francis Crick [40] which shows the diagrams for the transfer of information.

Ideas on Protein Synthesis (Oct. 1956)

That is, we may be able to have

where the arrows show the transfer of information.

Figure 15: Central Dogma of Molecular Biology [40]

2.9.1 General Transfers

a) DNA replication

DNA replication is the process that a cell uses to copy DNA with the help of enzymes and proteins. In order to start DNA replication, the double stranded DNA helix must be opened. Helicases and single strand DNA binding proteins unwind the DNA into two single strands. After that, DNA polymerase III, I, ligase and primase proteins work together to build copies of template DNA (Fig 16).

b) Transcription

The process of creating a complementary RNA from a DNA template is called transcription (Fig. 17). In protein coding regions, the resulting complementary RNA copy is called mRNA. mRNA naturally exists in single stranded forms and acts as a template for

protein synthesis. The enzyme called RNA polymerase transcribes DNA to mRNA. After transcription, mRNA is moved to the ribosome where it is translated into protein.

Figure 16: DNA replication [41]

Transcription has main three steps:

I. Initiation: DNA is transcribed by the enzyme RNA polymerase. In this step RNA polymerase attaches to the DNA at a specific area called the promoter region by using specific nucleotides sequences. Promoters are the regions of the DNA that signal initiation of transcript.

II Chain elongation: RNA polymerase moves along the one strand of the DNA, the template strand, and creates mRNA.

III. Termination: When the RNA polymerase reaches the termination sequence, it releases the mRNA and detaches from the DNA.

c) Translation

Translation is the part of protein synthesis that produces a specific amino acid chain by decoding the mRNA generated by transcription (Fig. 18). It is performed by the ribosome which is a component of a cell.

Figure 17: Transcription [42]

Translation has main four steps:

I. Activation: Amino acids are attached to the tRNA.

II. Initiation: mRNA binds to the small subunit of the ribosome and the ribosome moves along the mRNA until it reads the start codon AUG. At that point in time, the large subunit of the ribosome attaches to allow starting of translation and start codon AUG binds with the tRNA that has anticodon UAC and the bound amino acid methionine.

III. Elongation: In elongation another tRNA attaches to the ribosome next to the start codon and binds a new amino acid to the first one to form polypeptide chain. The binding process repeats until a full polypeptide chain is formed according to the sequence of bases in the mRNA.

IV. Termination: When the ribosome reads a stop codon (UAA, UAG, UGA) on the mRNA, the completed protein is released from the final tRNA then the last tRNA and the mRNA are detached from the ribosome.

Figure 18: Translation [43]

2.9.2 Special Transfers

a) Reverse transcription

The transfer of information from RNA to DNA is reverse transcription. It occurs in retroviruses (RNA virus), such as HIV or retrotransposons (amplifying the genetic elements) and telomere (a region of repetitive DNA sequence at the end of a chromosome) synthesis.

b) RNA replication

Producing a new RNA from RNA is RNA replication. It is used to reproduce some viruses. These viruses can be double-stranded or single-stranded RNA [44]. Double-stranded RNA viruses make single-stranded RNA molecules from the double-stranded RNA molecules. Single-stranded RNA viruses are divided into two groups: negative-sense single-stranded RNA viruses and positive-sensed single-stranded RNA viruses. The RNA molecule of negative-sense viruses cannot be read directly to create proteins. First,

complementary RNA is created and used to produce viral proteins. RNA molecules of positive-sensed single stranded RNA viruses can be read directly for the synthesis of viral proteins.

c) Direct translation from DNA to Protein

Translation of proteins directly from DNA has been shown in cell-free systems. DNA is obtained from mammalian cells and added into an *Escherichia coli* cell-free system. Later antibiotics are added to the system and protein is produced directly from single stranded DNA [45].

2.10 Genetic Code

When the information is needed to make a protein, one strand of a DNA molecule is transcribed to mRNA. Later the mRNA is translated to a protein composed of amino acid molecules. Since there are only four bases in mRNA to code the 20 amino acids, more than one base must be used to specify an amino acid. Even using two bases to code all 20 amino acids is not enough ($4 \times 4 = 16$). Therefore, three bases are required to decode one amino acid. A single set of these three bases is called a codon and the set of all possible combinations of three bases called the genetic code first discovered by Marshall Warren Nirenberg [46] in 1968 (Table 4). There are 64 ($4 \times 4 \times 4$) different combinations or codons. Three of them are stop codons which gives a signal to terminate the amino acid chain being synthesized on the ribosome. The start codon is AUG. It also encodes the amino acid methionine. The rest of the amino acids are encoded by the each of the remaining sixty codons.

TABLE 4: Genetic code

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The following is an example to show how codons decoded from mRNA to amino acid chain. “*” denotes stop codons and the sequence is partial, where it assumes that the start codon already passed.

DNA: UUA ACA UGA AAG AUG ACA UAC GAU AGC GAU GAU CGA CGC

Leu Thr * Lys Met Thr Tyr Asp Ser Asp Asp Arg Arg

L T * K M T Y D S D D R R

2.11 Alternative Splicing

Alternative splicing is a process which increases the biodiversity of proteins by allowing multiple proteins to be created from one section of a DNA. It is also called differential splicing.

A gene is first transcribed into a pre-messenger RNA which is a copy of a specific section of DNA containing both introns and exons. After introns are removed from pre-

mRNA, exons join together by the alternative splicing process. During alternative splicing, exons of a gene can be spliced together in different ways to create different mRNAs from that gene. As a result, the proteins coded by alternatively spliced mRNAs contain different amino acid sequences and often function differently.

Alternative splicing events can be categorized into five main types:

- **Exon skipping:** An exon of a gene can be excluded from the produced mRNA.
- **Intron retention:** An intron can remain in the produced mRNA.
- **Mutually exclusive exons:** One of two exons is included in the produced mRNA, not both.
- **Alternative donor site:** An alternative 5' splice junction is used.
- **Alternative acceptor site:** An alternative 3' splice junction is used.

Figure 19: Alternative splicing types

2.12 Transcript

A single gene can produce multiple proteins by using alternative splicing to create different transcripts. As a result, transcripts of the same gene differ in the exons used to construct the corresponding mRNA. For example, table 5 provides the transcript list of the gene *CTLA4* and the proteins that transcripts produce. From the *CTLA4* gene, a cell can have five different transcripts and produce four different proteins.

TABLE 5: Transcripts of the Gene *CTLA4*

Name	Ensembl Transcript ID	Protein
CTLA4-001	ENST00000302823	223aa
CTLA4-005	ENST00000295854	174aa
CTLA4-002	ENST00000427473	137aa
CTLA4-004	ENST00000472206	79aa
CTLA4-003	ENST00000487393	No protein

2.13 Noncoding DNA

Noncoding DNA is the region of DNA that does not contain protein coding information [47]. Although initially these regions were thought to be non-functional and therefore called junk DNA, noncoding DNA does have important functions in the genome such as they control when, where and what level gene expressed via promoters, enhancer and silencer regions. These regions may also produce non-coding RNAs (miRNAs, lncRNAs, sncRNAs, ...) The proportion of the noncoding DNA within organisms varies. For example, in the human genome 98% or more of the genome is noncoding, whereas in prokaryotes only 20% or less of the genome is noncoding.

2.13.1 Promoter

Promoters are adjacent to a gene near the 5' end of the transcription initiation site. They indicate the start point of gene transcription and provide a binding site for RNA polymerase. RNA polymerase initializes the transcription by binding to the promoter sequence. Many eukaryotic genes have an A-T rich promoter sequence called a TATA box. This name comes from the preserved sequence most commonly observed TATAAA. During transcription a TATA binding protein binds to the TATA-box to unwind DNA and then bend it through 80°. About 24% of human genes have a TATA-box.

2.13.2 Enhancer

Enhancers are DNA sequences that increase the expression of a specific gene when bound by transcription factor proteins (activator) [48]. Enhancer sequences can be located in both forward and backward direction. They can be located in the intron of the gene being regulated or up to one million of base pairs away from the gene. The long distance between the enhancer and the gene does not affect the function of enhancers because of the folded and coiled chromosome structure of DNA. Spatially they can be located near the transcription start site in chromosome structure.

2.13.3 Silencers

Silencers are DNA sequences that block the expression of a specific gene when bound by proteins (repressor) [49]. They are mostly located in the upstream of the targeted gene. The distance ranges from 20 base pairs to 200 base pairs upstream of a gene. Some silencers can be also located in the intron or exon of the targeted gene or in the 3' UTR of mRNA.

CHAPTER 3

MICROARRAYS

Curiosity concerning the building blocks of life has led to the development of technologies that have had huge impact on the fields. In the 1970's Edward M. Southern developed the technique of Southern blotting to locate particular gene or DNA sample in the genome via gel electrophoresis and probe hybridization [50]. In 1977 Frederick Sanger developed the Sanger sequencing method to determine the sequence of DNA by a chain termination technique based on the use of dideoxynucleotides in addition to the normal nucleotides found in DNA [51]. In the same year, James Alwine and George Stark developed the northern blotting technique to detect specific mRNA in a sample [52]. In 1983 Kary Mullis developed the polymerase chain reaction (PCR) technique to allow exponential amplification of DNA segments (thousands to millions of copies of a particular DNA sequence) [53]. Shortly after, PCR became a fundamental technique used in many medical and biological laboratory protocols. In terms of gene level analysis, researchers were using all these methods to work on one specific gene product per experiment. In 1995 Schena et al. published the first paper about DNA microarray technology which allows parallel analysis and investigation of thousands gene products using known sequence data that is complementary to the gene of interest [54]. More recent advances allow scientists to investigate the whole genome with or without prior information via next generation sequencing.

3.1 DNA Microarray

A DNA microarray (DNA chip or biochip) is a technology used to identify and measure the level of mRNA present in prokaryotic and eukaryotic DNA. It is a solid surface usually a glass, microscope slide or a silicon chip with fixed locations called *cells*, *spots* or *features* [55]. The spots contain millions of identical selected oligonucleotides called *probes* (Fig. 20).

Figure 20: Microarray features and probes

Each probe corresponds to a fragment of genomic DNA, cDNAs, PCR products or chemically synthesized oligonucleotides of up to 800 bases. These oligonucleotides are complementary to a gene of interest such as transcripts or exons. In microarrays, one or more oligonucleotides group together and represent a gene of interest. Some microarray platforms use microscopic beads in place of solid surfaces.

Microarrays have been used to identify relative gene expression to learn about cell functions, cell differentiation, genotyping, single nucleotide polymorphisms (SNPs), effect of treatment over diseases, changes of a particular gene involved in a disease and polymorphisms or mutations within a population.

3.1.1 Hybridization

Nucleic acid hybridization is the basis of microarrays [56]. It is the process in which similar single stranded nucleic acid sequences interact and form hydrogen bonds between complementary bases adenine (A) - thymine (T) and guanine (G) - cytosine (C). It can happen between DNA/DNA, RNA/RNA, DNA/RNA and short oligonucleotides.

Through nucleic acid hybridization, the degree of similarity between two nucleic acid strands can be measured and the complementary sequences of interest (targets) can be identified. In a microarray experiment, hybridization happens between probes and complementary target mRNAs which were obtained from test samples.

3.1.2 DNA Microarray Experiment

A microarray experiment starts with sample extraction from two different subject cells to be compared such as treated-untreated, healthy-diseased, and mutant-wild type. Subsequently samples are labeled. The resulting labeled samples are called *targets*. Once samples are prepared they are diffused over the microarray. There are two major types of DNA microarrays: one-color and two-color. When one-color microarray is used, samples are labeled with a single dye (such as phycoerythrin, cyanine-3 (Cy3), cyanine-5 (Cy5) or biotin) and hybridized to separate microarrays. When two-color microarrays are used, samples are labeled with two different fluorescent dyes and hybridized together on a single microarray. To remove unbound targets, the array is washed after hybridization.

One-color microarrays are also called single-channel microarrays. Since only a single dye is used and each sample hybridized to different microarrays, the obtained data

gives the estimation of absolute value of each gene's expression for each sample. Some of the benefits of using one-color system are:

- Comparing the absolute values of gene expression between different experiments is easier even when the experiments are done in different times and locations.
- An abnormal sample cannot affect the other samples' raw data because each sample is hybridized to a separate microarray.
- When the experiment becomes more complicated such as effect of treatment over time, one-color microarray is usually advisable.

Figure 21: One-color Affymetrix® GeneChips® microarray experiment diagram

In a two-color microarray, samples hybridize on the same microarray therefore it estimates the gene expression concentration ratios (each gene up or down regulated). Fig.

22 shows a schematic for a two-color microarray experiment that is designed to compare experimental sample (cancer cell) representing the expression pattern of genes in a specific set of conditions with a control sample (normal cell) representing all the genes that are expressed in the cells to be analyzed.

Figure 22: A typical two-color microarray experiment diagram

Multiple steps are involved in a two color microarray experiment.

1. Each experiment starts with isolation of mRNAs that represent the amount of genes expressed at the time of sample collection from the experimental sample. For example,

- cancer cells and a control sample (normal cells). The success of the experiment depends on the quality of the extracted mRNAs.
2. Next, the extracted mRNAs are converted into complementary DNA (cDNA) with the help of a reverse-transcriptase enzyme and labeled with a different fluorescent cyanine dyes (Cy3 and Cy5) to track the cDNAs coming from different samples.
  3. The labeled experiment and control cDNA are mixed together, and then purified. After purification, the mixed labeled cDNA is hybridized to microarray. The microarray is washed to remove any labeled cDNA that did not hybridize on the microarray and heated to reduce cross hybridization. Each labeled cDNA only binds to its complementary target sequence on the microarray.
  4. In the final step, an image of the microarray surface is produced by scanning to determine how many labeled cDNA probes bound to target spots in the microarray. In this experiment, red spots on the microarray represent genes upregulated compared to normal samples, green spots represent genes that are downregulated in the cancer sample, and yellow spots represents genes that active in both cancer and normal samples. The upregulated and downregulated genes can be further investigated with data mining techniques such as clustering.

In two-color microarray experiments dye bias may affect the identification of differentially expressed (up-regulated) genes and increase the false positive and negative results. The most common approach to correct dye bias is dye swap design. In dye swap design, the initial experiment and control samples are labeled with Cy3 and Cy5 and then hybridized on a chip. Later on, the process is repeated by switching the dyes of the two

samples. The results of two samples are averaged and used to identify differentially expressed genes.

3.1.3 Experimental Variations in Microarray Platforms

The purpose of a microarray experiment is measuring changes of mRNA levels between different state of interests such as cancer cell versus non-cancer cell to detect significant changes between states by applying analysis algorithms. But a detected significant change may not always be real. The significant change might be result of a random change, systematic bias in the biology, study, or samples [57] which is explained by computer science originated principal “garbage in, garbage out” that the quality of the analysis output received from a biological experiment depends on the quality of the experimental data. Thus, it is important to make sure that steps taken during the microarray experiment do not introduce deviations over the mRNA changes originated by the state changes of interests [58]. For example, biologist must make sure samples are not contaminated, and signals are not coming from other forms of RNAs included in the samples addition to mRNAs of actual interest of samples. This can be performed by biologist having control over all stages of experiments and analysis, using replicates and cooperating their biological knowledge with the data they are studying via microarrays. They can use mean, median standard deviations values to detect outliers, scatter plots and histograms to observe what is accruing in a microarray.

3.1.4 Types of the DNA Microarrays Based on Technology

Microarray fabrication varies based on the probe type, solid surface used, probe addressing method and target detection [59]. Different technologies are used to fabricate

microarrays.

Printed Microarrays: In the first microarrays chemically, synthesized oligonucleotides were printed or spotted onto very fine solid surface usually a glass via either noncontact or contact printing. In noncontact printing, the probe liquid droplets are applied onto microarray surface by the same technology used for computer printers. In contact printing, print pins are used to apply probes directly into a specific spot on the surface. Probes in printed microarrays are either double-stranded DNA (dsDNA) or oligonucleotide. dsDNAs range from 200 to 800 bases. Oligonucleotides range from 25 to 80 bases.

In Situ-Synthesized Oligonucleotide Microarrays: In situ-synthesized arrays are high-density DNA microarrays. The solid surface used for in situ-synthesized microarrays is typically a quartz wafer or glass. Probes in the arrays are chosen from oligonucleotides. Oligonucleotides are directly synthesized on the microarray surface. The manufactured in-situ-synthesized microarrays are Agilent, Roche NimbleGen and Affymetrix® GeneChips®. Agilent uses a non-contact industrial inkjet printing process to spot oligonucleotides onto specially-prepared glass slides. Agilent probe length is 60 bases and most genes represented by single probe rarely by couple of probes. Agilent technology places all four nucleosides simultaneously onto glass slide in repeated cycles of base by base printing, requiring only one synthesis cycle per layer which leads to having longer and more specific probes. Roche NimbleGen uses maskless photo-mediated synthesis. Probe length ranges from 60 to 100 bases. Currently Roche NimbleGen arrays are not available. Both Agilent and Roche NimbleGen allow multicolor experiment.

Alternative to Agilent and Roche NimbleGen arrays is Affymetrix® GeneChips® which is the most widely used array. Affymetrix® GeneChips® use semiconductor-based photolithography to construct individual probes of length 25 bases. Different than Agilent, multiple probes represent a gene. Affymetrix® develops and commercializes variety of arrays for plants, animals and microorganisms. In addition to predesigned arrays, they provide MyGeneChip™ service where researchers can design their own arrays for specialized studies. Data formats used by Affymetrix® are very well standardized. They also provide necessary experiment reagents, tools and software to obtain and analyze the results. Affymetrix GeneChips® are limited to one color based on a biotin labeling.

3.2 Affymetrix® GeneChips®

Affymetrix® GeneChips® are the most commonly used prefabricated arrays for gene expression analysis. In Affymetrix® GeneChips® 11-20 probes, each of them 25 nucleotides length, form a probe set and represent transcript variants of a gene. Probes are chosen from the region of a gene that has the least similar nucleotide sequence to other genes [56] and placed onto a silica wafer substrate via a photolithography technique as probe pairs (Fig. 23). Each probe pair has one perfect match (PM) probe and one mismatched probe (MM). In the most recent Affymetrix® GeneChips®, the MM probes are not used therefore they do not have probe pairs.

Figure 23: Affymetrix® GeneChips® microarray design

PM probes have an exact complementary sequence to the transcript of a gene based on a reference genome (Fig. 24). Each MM probe differs from the perfect match probe by a complementary base located in the middle of a probe sequence (13th base). PM probes help to measure the expression level of gene transcripts, while MM probes help to inspect PM probe and detect cross hybridization events and background signals.

Figure 24: Affymetrix® GeneChips® probe selection and perfect match (PM), mismatch (MM) probe sets

3.2.1 Array Design

The successful Affymetrix® GeneChip® array design requires selection of unique multiple probes for each transcript. Affymetrix® GeneChip® probe set design involves collecting sequences, annotating sequences via clustering, selecting a single sequence that represents the cluster and tiling this sequence into 25 base probes.

Probes are selected from consensus or exemplar sequences. The strategy for probe selection regions varies according to the purpose of the array. Some arrays may examine the gene expression in the level of exons and isoforms others may focus on polyadenylation sites. In the HG-U133 arrays more than one probe selection regions were selected in order to cover alternative polyadenylation sites. To decide the probe selection regions, potential transcript ends were identified by any of the following criteria (Fig. 25):

- The 3' end of a potential full length member sequence (RefSeq and complete CDS mRNA sequences).
- A set of eight or more ESTs ending at the same position.
- 3' end of the consensus sequence, a sequence obtained from the sequence of the cluster members.

Later on, probes were selected from the 600 bases most nearest to the 3' end of exemplar or consensus sequence (Fig. 25). In the case of selecting probes within the putative transcript ends, the exemplar mRNA sequence is used. For all other transcript ends the consensus sequence is used. When the orientation of cluster is unknown or uncertain the probes were selected from both ends of the sequence.

Figure 25: Affymetrix® GeneChips® Human Genome U133 multiple probe selection regions (Figure adopted from Array Design for the GeneChip® Human Genome U133 set technical note [60])

Figure 26: GeneChips® Human Genome U133 arrays transcript, consensus sequence, and probes relationship

In the human Exon 1.0 arrays probes were selected from exons of the consensus sequence (Fig. 26).

Due to multiple potential probe selection regions, some heuristic rules are applied to give priority for well annotated and strongly supported regions so they can be presented in the array. The regions represent the mRNAs, annotated as containing the complete coding sequence and 3' untranslated sequences. Evidence of polyadenylation, the size of cluster, orientation and genomic mapping were used for EST-only clusters prioritization decisions.

The heuristic probe selection strategies were intended to select probes unique to a single transcript or common among a small set of transcripts variants, but the hybridization characteristic of probes was not considered. With the designed technique for HG-U133, the probe binding characteristic was examined with multiple linear regression models based on a nucleic acid duplex formation thermodynamic model. The technique predicts the probe binding affinity and linearity according to the varying target concentrations.

After probes were selected, they were grouped together to form a probe set. The HG-U133 set contains 11 probe pairs per probe set. The older version has 16-20 probes per probe set. In human genome U133 Plus 2.0 arrays, the probe set names were marked via suffixes (Fig. 27). If all probes were perfectly matched to transcripts of same gene, they were suffixed with “_a”. If all probes were perfectly matched to multiple transcripts of different genes, they were suffixed with “_s”. If probes were highly similar or identical to other sequences, they were suffixed with “_x”. The probe sets with probes match to single transcript were not suffixed.

Different probe set types are indicated by suffixes to the probe set name. Unique probe sets are predicted to perfectly match only a single transcript. Gene probe sets, with an "_a" suffix, are predicted to only perfectly match transcripts from the same gene. Common probe sets, with a "_s" suffix, are predicted to perfectly match multiple transcripts, which may be from different genes. Probe sets that have a "_x" suffix are not shown here but are described in the text.

Figure 27: GeneChips® Human Genome U133 Plus-2 probe set marking (Figure adopted from Design and Performance of the GeneChip® Human Genome U133 Plus 2.0 and Human Genome U133A 2.0 Arrays technical note [61])

3.2.2 Attaching Probes

Affymetrix® uses a photolithography technique to attach probes into spots. It is a similar technique used for silicon computer chips where removal and positioning of silicon material on the chip is controlled by selectively exposing light with the help of a mask. Affymetrix® GeneChip® arrays photochemically modified synthetic linkers attached to the array surface to block the nucleic acid addition until deprotected by a light. Ultraviolet light is exposed to the array through the mask. Each mask is produced with windows that direct the order of nucleotide addition by either transmitting or blocking the ultraviolet light. When the mask has an open window at a specific spot, the spot gets deprotected and becomes available for specific nucleotide addition (Fig 28).

3.3 Data Files

3.3.1 Affymetrix GeneChip® Data Files

Affymetrix® GeneChip® technology uses special data formats to store GeneChip® array information and the experimental results [62]. In this research, we obtained the performed experimental results as CEL files and modify the original Chip Description Files (CDFs) to create custom CDFs according to the current genomic data.

Figure 28: Affymetrix GeneChip microarray probe attaching via masked photolithography (Figure adopted from Basic Concepts of Microarrays and Potential Applications in Clinical Microbiology [59])

Some of the important files are:

- a) **EXP File:** The EXP file contains information about experimental conditions and protocols. It has three sections: sample information (chip type, sample type, operator), fluidics (protocol, station, hybridization date), and scanner (pixel size, scanner type, scan date).

b) DAT File: The DAT file contains the pixel intensity values of the image produced by the Affymetrix® scanner. It is a binary file and contains two sections: header and pixel intensity data. It starts with header section and followed by pixel intensity data section. Header has the dimension of the image (pixel number, coordinates of the grid, number of rows etc.), the scanning conditions (temperature, scan speed, laser power, etc.) and intensity related values (mean pixel value, minimum and maximum pixel value etc.) The pixel intensity data has the intensity values of each row stored as 16-bit unsigned integer values.

```
Affymetrix GeneChip Experiment Information
Version 1

[Sample Info]
Chip Type      Chicken
Chip Lot       No
Operator       grt
Sample Type    chicken spleen
Description
Project AD1
Comments
Solution Type
Solution Lot

[Fluidics]
Protocol       EukGE-WS2v5_450
```

Figure 29: Part of the ASCII EXP file of chicken genome array supplied by Affymetrix®
in the Sample Data for GeneChip® Arrays

c) CEL File: The CEL file contains the calculated intensity values of each probe based on the pixel intensities stored in a DAT file. It also provides the standard deviation of intensities, the number of pixels used for the intensity calculation, an outlier flag and a

user defined flag to indicate the spots that need to be left out during the analysis. It can be either in ASCII text format (Version 3) or in binary format (version 4).

d) Chip Description Files (CDF): The CDF file stores the layout information for an Affymetrix® GeneChip® array. Some of the important data described in CDF are: which probes belong to which probe set, spot coordinates of each probe, name of probe sets and PM-MM probe pairs (Fig. 31). It is used to read the intensity values of probe sets located in the CEL file to detect the expression level of genes or transcript variants of a gene. There are two versions of each CDF: an ASCII text format and XDA format.

```
[CEL]
Version=3

[HEADER]
Cols=984
Rows=984
TotalX=984
TotalY=984
OffsetX=0
OffsetY=0
GridCornerUL=219 193
GridCornerUR=7143 218
GridCornerLR=7118 7151
GridCornerLL=194 7125
Axis-invertX=0
AxisInvertY=0
swapXY=0
DatHeader=[13..20196] AD1chicken151:CLS=7331 RWS=7331 XIN=1 YIN=1 VE=30 2.0 03/14/05 12
Algorithm=Percentile
AlgorithmParameters=Percentile:75;CellMargin:2;OutlierHigh:1.500;OutlierLow:1.004;AlgVersion:6.0;

[INTENSITY]
NumberCells=968256
CellHeader=X Y MEAN STDV NPIXELS
0 0 100.0 21.0 25
1 0 9571.0 1160.7 25
2 0 164.0 34.1 25
3 0 9923.0 1607.1 25
4 0 138.0 29.0 25
5 0 79.0 16.2 25
6 0 9581.0 1383.0 25
7 0 123.0 29.0 25
8 0 9512.0 819.8 25
9 0 125.0 23.4 25
```

Figure 30: Part of the ASCII CEL file of chicken genome array supplied by Affymetrix®
in the Sample Data for GeneChip® Arrays

Figure 31: Part of the ASCII HG-U133_Plus_2.cdf file

- e) **The Gene Information (GIN) File:** The GIN file contains genomic information of a specific array such as the gene names associated with each probe set.

```
Version=2
Arrays=HG-U133_Plus_2
NURLS=0
Index Identifier Identifier Source Probe Set Name Description Group Sequence Type repAccession GeneName
1 AFFX-BioB-3_at E. coli /GEN=bioB /DB_XREF=gb:J04423.1 /NOTE=SIF corresponding to nucleotides 2755-3052 of
2 AFFX-BioB-5_at E. coli /GEN=bioB /DB_XREF=gb:J04423.1 /NOTE=SIF corresponding to nucleotides 2032-2305 of
3 AFFX-BioB-M_at E. coli /GEN=bioB /DB_XREF=gb:J04423.1 /NOTE=SIF corresponding to nucleotides 2482-2739 of
4 AFFX-BioC-3_at E. coli /GEN=bioc /DB_XREF=gb:J04423.1 /NOTE=SIF corresponding to nucleotides 4626-4878 of
5 AFFX-BioC-5_at E. coli /GEN=bioc /DB_XREF=gb:J04423.1 /NOTE=SIF corresponding to nucleotides 4218-4566 of
6 AFFX-BioDn-3_at E. coli /GEN=bioc /DB_XREF=gb:J04423.1 /NOTE=SIF corresponding to nucleotides 5286-5570 of
7 AFFX-BioDn-5_at E. coli /GEN=bioc /DB_XREF=gb:J04423.1 /NOTE=SIF corresponding to nucleotides 4980-5256 of
8 AFFX-CreX-3_at Bacteriophage P1 /GEN=cre /DB_XREF=gb:X03453.1 /NOTE=SIF corresponding to nucleotides 1089-
9 AFFX-CreX-5_at Bacteriophage P1 /GEN=cre /DB_XREF=gb:X03453.1 /NOTE=SIF corresponding to nucleotides 513-1
10 AFFX-DapX-3_at B. subtilis /GEN=jojG /DB_XREF=gb:L38424.1 /NOTE=SIF corresponding to nucleotides 2684-3130
11 AFFX-DapX-5_at B. subtilis /GEN=dapB /DB_XREF=gb:L38424.1 /NOTE=SIF corresponding to nucleotides 1428-1926
12 AFFX-DapX-M_at B. subtilis /GEN=dapB, jojF /DB_XREF=gb:L38424.1 /NOTE=SIF corresponding to nucleotides 202
13 AFFX-HSAC07/X00351_3_at H. sapiens /GEN=beta-actin /DB_XREF=gb:X00351.1 /NOTE=SIF corresponding to nucleoti
14 AFFX-HSAC07/X00351_5_at H. sapiens /GEN=beta-actin /DB_XREF=gb:X00351.1 /NOTE=SIF corresponding to nucleoti
15 AFFX-HSAC07/X00351_M_at H. sapiens /GEN=beta-actin /DB_XREF=gb:X00351.1 /NOTE=SIF corresponding to nucleoti
16 AFFX-HUMGAPDH/M33197_3_at H. sapiens /GEN=GAPD /DB_XREF=gb:M33197.1 /NOTE=SIF corresponding to nucleotide
17 AFFX-HUMGAPDH/M33197_5_at H. sapiens /GEN=GAPD /DB_XREF=gb:M33197.1 /NOTE=SIF corresponding to nucleotide
18 AFFX-HUMGAPDH/M33197_M_at H. sapiens /GEN=GAPD /DB_XREF=gb:M33197.1 /NOTE=SIF corresponding to nucleotide
19 AFFX-HUMISGF3A/M97935_3_at H. sapiens /GEN=ISGF-3 /DB_XREF=gb:M97935.1 /NOTE=SIF corresponding to nucleoti
20 AFFX-HUMISGF3A/M97935_5_at H. sapiens /GEN=ISGF-3 /DB_XREF=gb:M97935.1 /NOTE=SIF corresponding to nucleoti
21 AFFX-HUMISGF3A/M97935_MA_at H. sapiens /GEN=ISGF-3 /DB_XREF=gb:M97935.1 /NOTE=SIF corresponding to nucleoti
22 AFFX-HUMISGF3A/M97935_MB_at H. sapiens /GEN=ISGF-3 /DB_XREF=gb:M97935.1 /NOTE=SIF corresponding to nucleoti
23 AFFX-HUMRGE/M10098_3_at H. sapiens /GEN=18S rRNA /DB_XREF=gb:M10098.1 /NOTE=SIF corresponding to nucleotide
24 AFFX-HUMRGE/M10098_5_at H. sapiens /GEN=18S rRNA /DB_XREF=gb:M10098.1 /NOTE=SIF corresponding to nucleotide
25 AFFX-HUMRGE/M10098_M_at H. sapiens /GEN=18S rRNA /DB_XREF=gb:M10098.1 /NOTE=SIF corresponding to nucleotide
```

Figure 32: Part of the ASCII HG-U133_Plus_2.gin file

- f) **Probe Set Information (PSI) Files:** Probe set names and the number of probe pairs in a probe set. It uses ASCII text format. The first line of the file has the number of probe

set names with “#Probe Sets” tag, the rest of the file contains the probe set names and number of probe pairs in the probe set.

```
#Probe Sets: 54675
1  AFFX-BioB-5_at  20
2  AFFX-BioB-M_at  20
3  AFFX-BioB-3_at  20
4  AFFX-BioC-5_at  20
5  AFFX-BioC-3_at  20
6  AFFX-BioDn-5_at 20
7  AFFX-BioDn-3_at 20
8  AFFX-CreX-5_at  20
9  AFFX-CreX-3_at  20
10 AFFX-DapX-5_at  20
```

Figure 33: Part of the ASCII HG-U133_Plus_2.PSI file

3.3.2 FASTA File Format

FASTA format is used to describe nucleotide or peptide sequences in a text-based file [63]. A sequence in FASTA format has two parts: single line description and sequence data (Fig. 34). The description line starts with a greater-than (>) symbol. The word following > is the identifier of the sequence, and the rest of the description line is optional. The sequence data has the nucleotides or amino acids represented by single-letter codes. Simple FASTA structure makes parsing easier for text-processing programs.

```
>probe:HG-U133_Plus_2:1007_s_at:718:317; Interrogation_Position=3330; Antisense;
CACCCAGCTGGTCCTGTGGATGGGA
>probe:HG-U133_Plus_2:1007_s_at:1105:483; Interrogation_Position=3443; Antisense;
GCCCCACTGGACAACACTGATTCCT
>probe:HG-U133_Plus_2:1007_s_at:584:901; Interrogation_Position=3512; Antisense;
TGGACCCCACTGGCTGAGAATCTGG
>probe:HG-U133_Plus_2:1007_s_at:192:205; Interrogation_Position=3563; Antisense;
AAATGTTTCCTTGTGCCTGCTCCTG
>probe:HG-U133_Plus_2:1007_s_at:844:979; Interrogation_Position=3570; Antisense;
TCCTTGTGCCTGCTCCTGTACTTGT
```

Figure 34: A part of the Affymetrix® GeneChip® HG-133 Plus 2 FASTA file

3.3.3 General/Gene Transfer Format

General/Gene Transfer Format (GTF) is a tab-delimited text based file format which is used to represent a gene structure information. GTF is an extended version of the general feature format (GFF). It has additional columns specific to gene information. The columns in a GTF are:

- seqname : Name of the sequence usually chromosome, contig or scaffold id.
- source : Name of a data source. It can be a program name, project name etc.
- feature : Feature type name: Gene, transcript, exon, start_codon, stop_codon, CDS.
- start : Start position of the feature.
- end : End position of the feature.
- score : A confidence degree in the feature existence and coordinates.
- strand : Direction of the feature + is for forward, - is for reverse.
- frame : 0, 1 or 2. 0 means that feature begins with whole codon, 1 means that before the first whole codon, there is one extra base, 2 means that there are two extra bases before the first whole codon.
- attributes : Semicolon separated additional information about a feature.
- comments : Comments begin with hash (#).

The following is an example gene structure with nine exons and two transcripts taken from Ensembl gene annotation GTF file for human build 38.

```

#!genome-build GRCh38

1 Havana gene 11869 14409 . + . gene_id "ENSG00000223972";

1 Havana transcript 11869 14409 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id
"ENST00000456328";...

1 Havana exon 11869 12227 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id
"ENST00000456328";...

1 Havana exon 12613 12721 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id
"ENST00000456328";...; exon_id "ENSE00003582793";

1 Havana exon 13221 14409 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id
"ENST00000456328";...; exon_id "ENSE00002312635"; exon_version "1";

1 Havana transcript 12010 13670 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id
"ENST00000450305";...

1 Havana exon 12010 12057 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id
"ENST00000450305"; ...; exon_id "ENSE00001948541"; exon_version "1";

1 Havana exon 12179 12227 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id
"ENST00000450305"; ...; exon_id "ENSE00001671638"; exon_version "2";

1 Havana exon 12613 12697 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id
"ENST00000450305";...; exon_id "ENSE00001758273"; exon_version "2";

1 Havana exon 12975 13052 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id
"ENST00000450305";...; exon_id "ENSE00001799933"; exon_version "2";

1 Havana exon 13221 13374 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id
"ENST00000450305";...; exon_id "ENSE00001746346"; exon_version "2";

1 Havana exon 13453 13670 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id
"ENST00000450305";...; exon_id "ENSE00001863096"; exon_version "1";

```

Figure 35: A part of the Ensembl gene annotation GTF file for human build 38

CHAPTER 4

CUSTOM CHIP DESCRIPTION FILES (CDF)

In Affymetrix® GeneChip® technologies, probe sets are formed from eleven to twenty short oligonucleotide sequences called probes (see Section 3.2). After the hybridization of specific samples is performed with an Affymetrix® GeneChip®, the obtained intensity level of a probe set is interpreted as a gene expression level of the specific gene. Therefore, more accurate selection of probes leads to more accurate biological interpretation of experimental results.

Given a particular Affymetrix® GeneChip® platform, the design of the probes is fixed. However, the methods in which the resulting experiments can be analyzed are dynamic in nature due to the ability to annotate and group probes into uniquely defined groupings. This is particularly important given that there are 97,015 data series publicly available in the GEO [64] repository as of April, 2018. 4,873 data series sharing the exact same chip design which is Affymetrix® GeneChip® Human Genome U133 Plus 2.0. Each dataset consists of reassembled original submitter supplied records and curated data for biologically comparable samples. One way in which the analysis can dynamically change is by correcting the mapping between probe sets and genes by creating custom CDFs (Chip Description File) to arrange which probes belong to which probe set based on the latest genomic information or specific annotations of interest. Each analysis can improve the

results of the microarray experiment or make them more relevant to annotations specific to particular analysis.

This chapter mainly focuses on the issues of microarrays and previously suggested solutions to address these issues.

4.1 Limitations of Microarrays

Since GeneChip® designs are by their nature fixed, the selection of probes relied on earlier genome assemblies and annotation available at that time. Due to rapid progress of genome sequencing and annotation, significant differences can arise between the earlier and the current genome databases. The current databases are likely to be more accurate than the earlier ones, contain more annotated coding and non-coding genes, and reveal more information about alternative splicing of genes. Table 6 shows the original sequence numbers of some source databases used for Human U133 Plus 2.0 Affymetrix® GeneChip® at the time it was created and the latest ones. It is obvious there are significant differences that should be taken into account, making it necessary to update probe-gene mapping according to current databases. Affymetrix® does attempt to provide compatibility between genomic changes by updating the links between probe sets and their corresponding genes via NetAffx™ (<http://www.affymetrix.com>) [65]. NetAffx™ derives the functional and descriptive annotations of representative sequence from current releases of the UniGene [66], LocusLink [67] and Homologene [68] databases [69]. Representative sequence stands for the sequence that was represented by a probe set in the array. When a representative sequence could not be found in the databases, another representative sequence is assigned to probe set based on the originally assigned UniGene [66] cluster. But NetAffx™ does not provide a solution for problems caused by individual probes.

TABLE 6: The release date and the number of human sequences for UniGene, dbEST, Ensembl and RefSeq at the time of Affymetrix® HG-U133 design and the current versions

	Original version		Latest Version	
	Release Date	Human Sequences	Release Date	Human Sequences
UniGene	April 20, 2001(#133)	2.688.626	March 9, 2013(#236)	7,000,789
dbEST	April 28, 2001	3.471.886	Jan 1, 2013(#130101)	8704,790
Ensembl	July 2001(#1)	NA	Apr 5, 2018(92)	3,609,003,417
RefSeq	April 2001	12.716	Mar 13, 2018(87)	NA

Potential issues with probes in a probe set are:

- I. Probes in a probe set that no longer align to a genomic location. In the HG-U133 Plus 2 array a total of 40,680 probes out of 603,158 (excluding quality control probes) do not have a perfect match to the human genome assembly 38.
- II. Probes in a probe set might not match any transcribed sequence or might be targeting different genes than the designated gene of a probe set. Table 7 illustrates these two types of problems. In the original Human U133 Plus 2 Affymetrix® GeneChip® probe 228543_AT has 22 probes, but only 11 of them match the RefSeq [70] entry NM_001164811. The rest of the probes do not match any RefSeq location. Probe 228544_S_AT was originally created to measure expression level of the gene *CSRP2BP*. But half of its probes map to the *PET117* gene. Both *CSRP2BP* and *PET117* are located in the chromosome 20 of forward strand and 5,051 bases of their sequence overlaps. If one uses this probe to measure expression level of *CSRP2BP*, it could reflect a cross hybridization with *PET117*. Probe 228543_AT has an even worse situation. Only one probe matches to *CSRP2BP*, 11 of them match to *PET117* and 10 of them do not match to any transcribed sequence. The expression level measured by probe 228543_AT does not reflect the integrated expression of the designated gene and therefore causes misinterpretation in quantification.

- III. Multiple probe sets often represent a gene. Often those probe sets might be coming from the same splice variant or different splice variants. As long as probes were not marked with an id specific to one splice variant, the expression level of probes cannot be used as an indicator for the specific splice variant. In Table 7, both probe 225432_S_AT and 233396_S_AT match to *CSRP2BP* but at the transcript level; half of their probes match to NR_028402 and rest of them match to NM_020536. The measured expression level is inconsistent, and does not indicate which splice variant changes in their expression level and is not as informative as the probe sets that matches to a single splice variant.
- IV. Different probe sets targeting different genes shares common probes. Such non-specificity causes ambiguity and cross hybridization. In HG-U133 Plus 2 array total 36,493 probes out of 603,158 (excludes quality control probes) perfectly map to multiple locations on the human genome assembly 38.
- V. Some of the probe sequences contain single nucleotide polymorphisms (SNPs). SNPs are the genetic variation between individuals of the same species [71]. A SNP represents a base on a specific DNA location which differs between individuals. Most complex diseases are associated with SNPs. For the accurate analysis of experimental results, SNPS located within probes need to be detected and analyzed carefully. Some researchers have paid attention to the effect of SNP presence in Affymetrix® probes. Benovoy et al. [72] concluded that SNPs located in probes cause false-positives in Affymetrix® Human Exon analyses and the position of a SNP within a probe sequence effects the binding of targets to the probes. Rouchka et al. [3, 71] designed a methodology to test the effect of SNPs that located in the 13th base of a probe sequence

to hybridization rate of mismatch probes in Affymetrix® HG-U133A arrays. They reported that less frequently, MM probes hybridization is greater than PM probes when the target probe has a SNP in the mismatch location.

TABLE 7: The alignment of PM probes taken from HG 133 plus-2 array PM to RefSeq database together with their corresponding gene symbol and the designated gene taken from gin file. The alignments were achieved via Absolute Gene ID Conversion Tool [73]

Designated Gene	Gene	RefSeq	Probe Set			
			228544 S AT	228543 AT	225432 S AT	233396 S AT
			CRP2BP			
	<i>PET117</i>	NM_001164811	11	11	-	-
		NR_028402	-	-	11	11
	<i>CSRP2BP</i>	NM_020536	11	1	11	11

4.2 Custom Designed CDF Files

Several researchers have called attention to probe-gene mapping problems [69, 74-84]. To address these issues, they created their own CDF files in order to redesign probe sets. These approaches have a similar work flow but differ in the data source used, the selected target level (gene or transcript), whether to create probe sets from scratch or redesigning the existing groups and sharing probes between probe sets. Table 8 gives a summary of the previous work in the area of custom CDF construction. A few of them offer public access to their custom CDF file.

TABLE 8: Summary of available custom CDFs done by several researchers

Paper	Source Database	Target Biomolecular Entities	Number of Probes Per Probe Set	Organisms	Available	Mapping Tool
Chalifa-Caspi et al.	GenBank, Refseq, Ensembl, GeneCards, UniGene	Gene	unknown	Human	No	BLAT
Laurent Gautier et al.	RefSeq	Gene/Transcripts/ALU	unknown	Human	No	Bioconductor Package matchprobes
Dai et al.	UniGene, RefSeq, DoTS, Entrez, ENSEMBL, AceView...	Gene/Transcripts/Exon	3	Human, Mouse, Rat, Zebrafish, ...	Yes	unknown
Harbig et al.	GenBank	Transcript	unknown	Human	No	blastn
Liu et al.	RefSeq, GenBank	Gene/Transcript	unknown	25 different Organisms	No	UCSC BLAT
Lu et al.	AceView	Transcript	4	Human	No	BLAT
Ferrari et al.	GeneAnnot, GeneCards	Gene	6-11	Human	Yes	unknown
Yu et al.	GenBank, RefSeq, Ensembl, IPI	Gene/Transcript	unknown	Human, Mouse, Rat, Bovine, Zebrafish	No	BLAST
Leeuw et al.	Entrez Gene	unknown	unknown	Human, Mouse	No	unknown
Ballester et al.	Ensembl	Transcript	NA	Human, Mouse, Rat, ...	Yes	Exonerate
Risueno et al.	ENSEMBL, RNAdb	Gene/Transcript/Exon/ncRNA	unknown	Human, Mouse, Rat	Yes	blastn
Yin et al.	Ensembl, RefSeq, GenBank, Biomart, ZFIN	Transcript	unknown	Zebrafish	No	Exonerate

Chalifa-Caspi et al. [69] worked on updating the mapping of HG-U95 arrays via a three step methodology without disbanding the original probe sets. First, they performed a pairwise alignment between probe sequences and obtained transcripts from GenBank [85], NCBI RefSeq [70] and Ensembl [86] via BLAT [87]. They assigned the probes to a transcript when probes were in the mRNA orientation and had fewer than two mismatches. If a probe did not match any transcript, they were then aligned to ESTs and their representative EST accession number was stored. In the second step, the transcript was mapped to GeneCard [88] genes or UniGene [66] clusters if a GeneCard gene was not available. In order to convert UniGene results to GeneCard, they found corresponding GeneCard genes of obtained LocusLink [67]/Ensembl [86] genes from UniGene [66]. If the result did not have an annotated LocusLink gene, their corresponding genomic locations obtained from UCSC [89] and used to create a link to a GeneCard entry via GeneLoc. Obtained ESTs were mapped to UniGene cluster. In the last step, the transcript and gene annotations were recorded and used to create probe set to GeneCard annotation. If a probe matched to transcripts of a GeneCard gene, this probe was marked with this gene. In addition to marking, they calculated a sensitivity score of the probe set by dividing the number of major marked genes to the total number of probes and specify by finding how many other genes maps to this probe set and with how many probes.

Laurent Gautier et al.[74] also worked on Affymetrix® Human GeneChip® HG-133A and HG-U95Av2. They aligned PM and MM probes to NCBI's human RefSeq database [70] via a string matching Bioconductor package called matchprobes [90]. Only PM alignments were accepted. They examined the probes that matched to multiple reference locations in detail. They divided them into four analyzing groups. The first group

included the probes that matched to the same gene's different alternative splice variants. They decided to keep them. The second group included the probes that matched to 300 to 600 different locations. When they study those probes, they discovered they were designed to match human ALU repeats. ALU elements are highly repetitive short (around 280 to 300 base pairs) DNA sequences [91]. Rather than removing those probes, they put all of them into one probe set and called it human ALU. The third group was for the probes that matched to closely related annotations. Since separation of these genes requires expert annotators, they simply ignored them. The last group contained the probes that matched to multiple dissimilar annotations. They also removed them to prevent misleading analysis. As a result of their study, they created open source software called *altcdfenvs* and made it available to everybody as a Bioconductor package along with the data they used. However, it is no longer publicly available.

The most effective (comprehensive) study for probe-annotation remapping was achieved by Dai et al [75]. Rather than focusing on one reference database or combining multiple ones to create one custom CDF, they worked on different databases and created specific custom CDFs for each database. They proposed two different filtering and regrouping frameworks; one for the UniGene database and one for RefSeq, Database of Transcribed Sequences (DoTS)(<http://www.cbil.upenn.edu/downloads/DoTS/>), Entrez gene [92], Ensembl [86] gene databases. According to their manuscript [93] some of the UniGene cluster sequences are unreliable and the strand direction is unknown. In order to get accurate UniGene based probe sets, special steps were required. First they started with sequence alignments of probes and UniGene sequences.

(I) UniGene sequences were aligned to the most recent genome assembly and genome alignments were saved. This operation helped them to clean unreliable sequences in the specific UniGene cluster.

(II) Individual probe sequences were aligned to the genome sequence.

- Only perfectly matched probes were accepted. They also required each probe to have one perfect match to the corresponding genome sequence. They believed that the ones aligned to the non-transcript regions and one transcript region could be kept but it is better to remove them for more reliable results.

(III) Individual probe sequences were aligned to UniGene and dbSNP to form new probe sets.

- A probe that is a perfect match to a genomic region also must align with mRNA/EST sequences located in the UniGene database. The only exception for this rule was exon-exon junction probes that had a perfect match to mRNA sequence in the same UniGene cluster. They added them to the probe set with the lowest index number.
- Probes that matched with multiple cDNA/EST sequences of different UniGene clusters were removed to make probes specific to only one UniGene cluster. They believed that the probes in a gene specific probe set should include the sequences that are common to all alternative splice variance of the gene so that the results were not be affected by the tissue type. Because the knowledge of alternative splicing genes was not complete, they decided to group all the probes that aligned to the same gene to create a gene specific probe set.

- In addition to gene specific probe sets, they created transcript and exon specific probe sets in order to study alternative splicing events. In addition, they created probe sets that target the 3' end of probes.
- If the targeted gene does not have a known mRNA/reference sequence, they required that all the probes must be in the same direction on the genome. If the probes in a probe set were perfectly matched to genomic regions in a different direction and there is no mRNA/reference sequence in the UniGene cluster to point out the transcription orientation, they divided them into two probe sets, one for each direction.
- They required all the probes, which represent one UniGene cluster, be aligned continuously on the genomic sequence in the same direction except when mRNA reference sequence of the targeted UniGene cluster aligns to the different locations. If the probes spread across different genome locations or chromosomes, the longest contiguous set of probes were used to represent the UniGene cluster.

The well defined orientation and sequence information in the RefSeq, DoTS, Entrez, and ENSEMBL databases let them to create new probe sets easier than the UniGene database probe sets.

(I) First, they mapped the probe sequences to the corresponding genomic sequence and target sequences. Only the probes that had one perfect match to the corresponding sequence were accepted.

(II) After alignment, they grouped the probes that mapped to the same target and had correct orientation. They ordered the probes based on their location on the exon. Probe sets that have fewer than three probes were removed. They created gene, transcript, exon and 3' probe sets.

Besides updating probe sets based on the latest databases, they removed the perfect match and mismatch probes that have an allele-specific base in the center of the sequence. This was done to reduce the noise caused by single nucleotide polymorphisms (SNPs) in different samples. These probe sets were called allele-independent.

For gene and transcript specific probe sets they paired database ids with GenBank accession numbers in order to supply more stable ids for probe sets.

Harbig et al.[76] worked on alternative mapping of Affymetrix® HG-133 Plus 2.0 arrays. Unlike the others, they mapped the target sequences of probes that were obtained from Affymetrix® rather than probe sequences. A target sequence is an exemplar region of a specific transcript. It has ≤ 600 bases and was used to select short oligonucleotide sequences (probes). Target sequences were mapped to GenBank records via blastn [94] with a word size of 28. They saved the human RefSeq matched results (mRNAs); if there were not any then they saved the matched human sequences (cDNAs). The sequences were saved in the local database together with the original probe sets names and the calculated score based on the type of sequence and alignment. Later the probe sequences were compared with the collected sequences from the first step. The sequences were scored based on how many bases of a probe in a probe set exactly matched to the sequence. The best average score across all the probes was selected and assigned as a primary

identification of the updated probe set. When the probe set had the same score for several sequences identified, the researchers marked those probe sets as detecting multiple transcripts. They kept the probe sets with low score values because they believed they also contribute. In addition, they compared probes to the most common 8 human ALU sequences and marked the ones matched to them. The probe sets with probes matched to a completely different sequence from the rest of probes was evaluated manually. If the probe set did not match to any sequence then it was compared to all GenBank sequences until a match was found. Eventually they remapped 37% of the Hg133 plus 2.0 probe sets. The website mentioned in the paper for methods and results are not available anymore.

Liu et al.[77] developed the AffyProbeMiner software to regroup probes and remap probes to the targets for different organisms. They used RefSeq transcripts and validated complete coding sequences in the GenBank as their target sequences. During the construction of complete coding sequence database, they obtained the sequences that included “complete CDS” and did not include “intronic transcript” or “BAC clone” in their record. And also, they obtained RefSeq mRNAs whose accession number starts with “NM_”. The obtained complete CDSs aligned to the genome, and the ones with less than 99% identity were removed from the database. After that, poly(A) tails were trimmed from complete CDSs. As a last step of database creation, they removed one of the duplicated records which was mapping to the same gene or one sequence covered by the other one with less than or equal to 1% dissimilarity. The UCSC BLAT [87] server was used to map probes sequences to created database. According to mapping results, they proposed four different probe set grouping methods. They are transcript consistent (each probe in a probe set maps to the same set of transcripts), transcript unique (each probe in a probe sets maps

to the one transcript), gene consistent (each probe in a probe sets maps to the transcript of the same set of genes) and gene unique (each probe in a probe sets maps to the transcript of the one gene). They supplied CDFs of 25 different organisms for the R Bioconductor environment, Affymetrix® GCOS and dChip. They claimed that providing a flexible tool to create a CDF file and using the BLAT server and Perl (They provided faster process than only using R) made their work better than others. Currently their tool, CDF files and web site are not publicly available. Their method remapped the 65.6% of probe sets on HG-U133A array.

Lu et al. [78] proposed a method that differs from others by the target database used. They used the human transcript database AceView [95]. The AceView database was formed from GenBank, dbEST and RefSeq mRNAs. Since it is a combined database, more transcripts are covered than other databases thereupon mapping between probes and mRNAs resulted with more matching. They matched a probe with a transcript when the number of identical bases between them is more than 21 (out of 25 length probe sequence). The probes with the same target(s) grouped together and composed one probe set. They made sure that probe sets never share probes. To uncover the effects of the number of probes per probe set over expression results, they created artificial data sets from the Affymetrix® U133A Latin Square data [96] and ran simulations over them. Latin square data is a test dataset provided by Affymetrix® to develop and compare analysis methods. It is formatted as Latin square; $n \times n$ array of n different symbols in which each symbol occurs only one in each row and in each column of the array. They analyzed the data with the specially designed CDF which contains probe sets of randomly selected probes with a variable number. i.e. for CDF d2, two random probes selected from each probe set.

Simulation results clarified that when the number of probes in a probe set increase, the reliability of expression measurements enhanced and the number of false positives decreased. There is a prominent difference between the probe set size smaller than four and bigger than three. Therefore, they only kept the probe sets with more than three probes. The web site provided in the paper is not available anymore. With their proposed method only 22% of original probes stayed without any change.

Ferrari et al. [79] worked on developing CDFs for Affymetrix® human arrays based on the GeneAnnot [69] database. GeneAnnot contains the relationship between the Affymetrix® probes and transcript sequences obtained from cDNAs, GenBank, RefSeq and Ensembl. The probes were matched to obtained transcripts from previously mentioned databases and then linked to GeneCards. Chalifa-Caspi et al. gave a more detailed explanation of GeneAnnot construction methods already explained in this chapter. In GeneAnnot, a gene was described with the HUGO gene nomenclature committee (HGNC) database, Entrez Gene and Ensembl. All individual probes linked GeneCards gene reached from GeneAnnot and the ones that linked to the same gene combined to create new probe set. They created two CDF files; one had at least six probes per probe set and other one had at least 11 probes per probe set. They analyze the same data with brain array CDFs, AceView CDFs and their CDFs to compare the obtained differentially expressed genes. They used robust multi-array average (RMA) [97] and significance analysis of microarrays (SAM) [98] methods in R environment for analyzing. Most of the time around 40% of the genes was commonly detected by all the CDFs. The highest similarity (on average 82%) was obtained when their CDF's results were compared with brain array [75] Entrez and RefSeq CDF results. Affymetrix® arrays can be obtained from

<http://genecards.weizmann.ac.il/geneannot/customcdf.shtml>. They also supply a search engine to get specific information of a probe set and gene.

Yu et al. [80] were only interested in the probe sets that target the protein coding transcript. For this reason in addition to GenBank, RefSeq and Ensembl, they used the IPI (International Protein Index) [99] database to connect mapped mRNAs to external protein accession numbers and get only the probes that target protein coding sequences. At first they got the original probe set gene mapping information from the Affymetrix® NetAffx™ [65] website. As mentioned in the paper their blast procedure based on the GeneAnnot approach developed by Chalifa-Caspi et al. (summarized previously). Later they connected the selected target mRNAs from blast filter to the protein IDs in IPI database. Connected protein IDs were used to create new probe sets. As a result some of the probes classified to four groups; transcript unique (all the probes mapped to one protein), gene unique (all the probes mapped to alternative spliced proteins of same gene) cross-hybridize (maps to proteins of multiple genes) and not categorized (not connected to any protein). The probes that did not map to any protein were eliminated. When the probe matched to multiple variants of the same gene they annotated it with gene level annotation. When probe matched to the single variant of a gene they annotated the probe in transcript level. At the end they covered the 85.4% of original Affymetrix® target genes and 68.8% probes were kept in the new CDF file.

Leeuw et al. [81] believed that probes removed during custom CDF creation might be important for researchers to study not yet well established genes and proposed a method called CDF-Merger. CDF-Merger aims to regain the lost information due to removed probes by creating Hybrid CDFs from Affymetrix® CDF, Affymetrix® annotation files,

Brainarray custom CDFs and NCBI Entrez Gene Info Files. First they customized the Affymetrix® CDF and Brainarray CDF. The rules of customizing Brainarray CDF are: (I) Redefine the probe-set ids (atd_Entrez) to be able to track (II) Put “_d” to the probe set name when probe set has probes that is shared with other probe sets. The rules of customizing Affymetrix® CDF are: (I) Keep the original Affymetrix® probe set names for tracking. (II) Discard the probe sets that contained more than two probes which was already used in Brainarray CDF. (III) If a probe set shares one or two probes with Brainarray probes then remove the probes from probe set and keep the rest the probes, add _1 or _2 to the probe set name. (IV) Check the Affymetrix® annotation file with NCBI Entrez GeneInfo. If a probe set can be annotated by more than one Entrez id then leave annotations of these probe sets empty. (V) Remove the annotations of probe sets already targeted in the Brainarray CDF. (VI) Rename all the probe set names with pattern “atm_Entrez_id”. (VII) Unite the probe sets that share the same Entrez id and mark them by putting _m to their name. After customization they combined the all probe sets into one CDF and created R environment CDF.

Ballester et al. [82] performed annotations for multiple organisms. They started updating by mapping PM probes to the corresponding genome via the exonerate alignment tool [100]. They discarded the probes that aligned to genome with more than one mismatch and probes that aligned to more than a hundred times on the genome. Later on they mapped probe sets to the Ensembl transcripts. If at least half of the probes in a probe set mapped to a specific transcript then they annotated the probe set with this transcript. The results can be reached from the Ensembl web site. From the BioMart search tool [101], one can get all the annotations for a specific array. In addition to web access, BioMart can be also accessed

via the R programming language with the Bioconductor package biomaRt. They also have a perl API to let users reach the data. The annotated GeneArrays are arrays from Affymetrix®, Illumina and CodeLink.

Risueno et al. [83] worked on updating probe mappings of Affymetrix® human, mouse and rat microarrays and shared the results, files and created tools from a web site called GATEplorer (Genomic and Transcriptomic Explorer). As in other studies they mapped the probe sequences to the target database and annotated them according to the mapping results. Mapping was performed with blastn. Target sequences of mRNAs and cDNAs were obtained from Ensembl. Only PMs were accepted. Different than other approaches, they mapped probes that were not mapped to the coding regions to the non-coding RNA (ncRNA) and created CDFs from the mapping results. They used Ensembl and RNAdb [102] as a source for ncRNAs. The CDFs are classified into four groups: probes to gene loci, to transcripts, to exon and to ncRNA. They can be downloaded as an R library. They supply txt files that show the mapping of probes to gene loci and transcripts with Ensembl id, gene name and symbol. One can search a specific gene and examine it with the visualization of loci, transcripts, exons and mapped probes on a chromosome. They also give a presentation of gene expression patterns obtained from GeneAtlas [62] based on probe sets for a specific gene. In addition to gene based search, probe set, genome coordinate, sequence of nucleotide and amino acids based searches can be performed from the web site.

Yin et al [84] adopt the Dai et al.'s proposed method with some main changes to remap Affymetrix® zebrafish genome arrays. To determine genome specific probes, first they mapped the probe sequences to genome version 7 via Exonerate. They removed the

probes that did not match the genome and transcripts or matched to multiple locations. To be able to detect cross exon boundaries, they kept the probes that had no matching to genome but matched the transcript. Later the probes were aligned to transcripts that were composed from Ensembl, RefSeq, GenBank, Biomart, and ZFIN [103]. They downloaded the transcriptome alignment coordinates and coding sequence coordinates from the UCSC genome browser and followed the procedure Exlink [104], proposed by the Ensembl group, to organize transcripts of different databases. The idea behind Exlink is clustering the transcripts under same gene when their coding regions are overlapped. During the mapping process, the probes that mapped to multiple genes, intergenic regions or introns of genes were removed. The probes that passed the filtering step were clustered based on the transcripts they mapped. One specific cluster was either representative of a transcript or multiple common transcripts of the same gene. Since the reverse complementary probes generates weaker signal than positive strand probes, they were grouped separately from positive strand probes. They also wanted to pay attention to the effect of the selected database over probe remapping by performing an experiment. They examined the one specific data set with probe mapping created based on Ensembl and probe mapping created based on a multiple source database, UCSC. The differently expressed genes detected by multiple databases were more comprehensive than Ensembl and the multiple source database was also able to detect more alternative splicing events than Ensembl. To strengthen their conclusion they performed real time PCR validation for a select number of genes. Only UCSC were able to detect PCR validated genes.

In our work we also redefine the mapping between probe sets and target genes by creating custom CDFs (Chip Description File) based on the latest genomic information. In

contrast to previous approaches we annotated probes in UTR and individual exon levels in addition to gene and transcript levels. Furthermore, we used interval alignment via Nested Containment List (NCList) rather than sequence alignment. We also provide flexibility for users to create custom CDFs for their specific annotations of interest.

CHAPTER 5

COORDINATE MAPPING BETWEEN PROBES AND ANNOTATIONS

Mapping between sequences, genomes and genome annotations is one of the main process of bioinformatics related works. For example, assembly of multiple sequences is used to construct the complete genome of an organism. Mapping between a query sequence and specific regions of an organism genome is used to identify annotations of the query sequence. Mapping used for comparative genomics query deals with finding common annotations between different genomes.

In bioinformatics, biological annotations in a genome such as genes, transcripts, UTRs, exons, introns and SNPs are modeled as intervals. For example, the human gene *GRIK4* is located on the chromosome 11 of human genome assembly 38, starts at 120,511,700 and ends at 120,988,904. It has twenty-one exons. One of exons whose Ensembl id is ENSE00003810695, starts at 120,511,887 and ends at 120,511,887. Therefore, finding annotation of a query sequence requires coordinate mapping. In our research, we used coordinate mapping to update mapping between probes and target genes by searching coordinates of a probe against coordinates of a genome.

Several structures have been applied to interval overlap searches in genomics. A BTree (MySQL database) with binning method has been used by the UCSC genome Browser [89]. The binning method partitions data into hierarchical bins and search the

intervals only in the matching bins. But it is not efficient for counting overlaps because all candidate bins need to be searched to find possible overlaps. R-tree indexing can also be applied for interval overlap searching. One of the variation of the general R-tree is interval trees. Interval trees are used by composing the tree from genomic coordinates of genes and searching coordinates of probes in the tree. Another faster and widely used way of performing interval overlap search is creating interval database representation of genomic coordinates via nested containment list (NCList) and searching probe coordinates in the NCList database. In our research, since the probes within a probe sets are short nucleic acid sequences, their genomic coordinates are not defined. We used the Bowtie alignment tool to obtain the genomic coordinates of probe sequences. It was also used to detect nonspecific probes. After obtaining probe coordinates, we used Nested Containment List (NCList) to perform mapping between probes and genomic annotations.

5.1 Bowtie

Bowtie (<http://bowtie-bio.sourceforge.net/index.shtml>) is a tool for aligning short read sequences to the large genomes [53]. It is fast and memory efficient. There are two main versions: bowtie 1 and bowtie 2. Bowtie 1 allows alignment of short reads to long reference genomes (e.g. mammalian). Bowtie 2 allows alignment of relatively long reads (100s of characters) to long reference genomes. It provides gapped, local and paired-end alignment modes.

Bowtie's reference genome is indexed with a Burrows-Wheeler transform so that its memory footprint becomes small. It uses an extended version of exact mapping called inexact matching to align short reads to an indexed genome. Backtracking and double indexing are two major extensions. It allows mismatches with high quality alignments

through the quality aware backtracking. Bowtie avoids excessive backtracking via double indexing.

In our work we used bowtie version 1.0.1 to align short probe sequences to the genome thereby the probes that no longer align to a genomic location or perfectly align to multiple genomic locations are removed and cross-hybridization effects are reduced. The following sections describe the algorithms used by Bowtie.

5.1.1 Burrows-Wheeler Transform

Burrows-Wheeler transform (BWT) is an algorithm developed for efficient data compression (Fig 36). It takes a character string and rearranges it. The result of the transformation contains the exact same characters that were started with, differing only in their ordering. It is useful for compression because it orders repeated characters next to each other and makes the data more compressible by algorithms such as run-length encoding [105]. Because it allows large texts to be searched efficiently with a small memory requirement, it is applied to sequence alignment in bioinformatics. The transformation is reversible.

Figure 36: A Burrows-Wheeler transform (BWT)

Without any additional data, BWT of a character string supplies information for:

- The number of occurrences of a specific character in the text.

Since the $BWT(T)$ is the permutation of the original text, we can count the number of occurrences of a character in the $BWT(T)$.

- Reconstructing the first column of the matrix M .

$BWT(T)$ has all the characters in the text. Sorting the $BWT(T)$ lexicographically will return the first column of the matrix M .

- Whether a specific two characters length substring is present or not in the original text.

The first property of Last-First (LF) Mapping formulated by Ferragina and Manzini [106] is used to search for a substring. $BWT(T)$ is ordered lexicographically and the rows that has the first character of the substring are used as index to look for the second character of the substring in the $BWT(T)$. If at least one of the marked rows has the second character then we can say that the original text has the substring.

- The second column of the matrix M .

It can be recovered by Last-First (LF) Mapping. From the last and the first columns, we can get the set of two consecutive characters in the original text. Since each row of the matrix M is cyclic shift, the same consecutive characters must be occurred between the first and the second column also. Based on this property, we can fill the second column. The same property is also used to recover the rest of the matrix M .

5.1.2 Inexact Matching

To map reads to the genome one must consider sequencing errors, mismatches, and genuine differences between the reference and query. But the exact matching does not handle these conditions. To make exact matching appropriate for genome mapping, bowtie

developed an algorithm called inexact matching. It uses backtracking to find alignments that satisfy the alignment policy which “allows a limited number of mismatches and prefers alignments where the sum of the quality values at all mismatched positions is low” [53].

The procedure of inexact matching is similar to the exact matching unless range does not become empty. When the range becomes empty, the algorithm might replace already matched query position by another base to allow mismatches between the reference and the query and then search resumes after the mismatch position. To substitute a base, it has to be consistent with the alignment policy and modified query has to be present in the text. When there are multiple candidates for substitution position, the position with a maximal quality is selected.

Backtracking was implemented via a stack data structure. The stack structure grows when a substitution candidate selected and shrinks when the substitution candidate is rejected. The alignment found via Bowtie might not be the best possible alignment in terms of number of mismatches and quality but the tool allows user to select best option to find the best possible alignment in terms of the number of mismatches. But it comes with a time cost.

5.1.3 Double Indexing

When more than one mismatch is allowed, a search might be too slow because of excessive backtracking. To avoid this, bowtie introduced a new technique called double indexing. Bowtie indexes both the genome (forward index) and reverse (character reverse) of the genome (mirror index). Searching the mirror index with reversed query is the same with normal search that progress left-to-right instead of right-to-left.

Figure 37: Exact matching versus inexact alignment. (figure adapted from [53] under Creative Commons Attribution Non-Commercial License)

5.2 Interval

An interval is a value pair of real numbers $[a, b]$ such that $a \leq b$. It represents the set $\{x \in \mathbb{R} : a \leq x \leq b\}$. a and b called the endpoints of the interval. The endpoints can be included or excluded from the set.

- When both endpoints are excluded from the set, it is called an open interval. They are represented as:

$$(a, b) =]a, b[= \{x \in \mathbb{R} \mid a < x < b\}$$

- When one endpoint is included in the set, it is called a half-open interval. They are represented as:

$$[a, b) = [a, b[= \{x \in \mathbb{R} \mid a \leq x < b\} \quad \text{OR} \quad (a, b] =]a, b] = \{x \in \mathbb{R} \mid a < x \leq b\}$$

- When both endpoints are included in the set, it is called a closed interval. They are represented as:

$$[a,b] = [a,b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$$

Between two intervals x and y , there can be three different relations.

- x and y overlap when $\text{low}(x) \leq \text{high}(y)$ and $\text{low}(y) \leq \text{high}(x)$.
- x is to the left of y when $\text{high}(x) \leq \text{low}(y)$.
- x is to the right of y when $\text{high}(y) \leq \text{low}(x)$.

5.3 Interval Tree

An interval tree is an augmented red-black tree [107]. It provides efficient access to set of intervals (Fig. 38).

A red black tree is a binary search tree with extra color information. The color makes the tree approximately balanced by controlling the node coloring for every path from a given node to any of its leaf. A red-black tree must satisfy the following properties:

1. “Every node is either red or black.
2. The root is black.
3. Every leaf (NIL) is black.
4. If a node is red, then both its children are black.
5. For each node, all paths from the node to descendant leaves contain the same number of black nodes.” [107]

Every node in an interval tree contains extra two fields; an interval defined by low and high value pair and the maximum value stored in the subtree of the node.

$$\text{max}[x] = \max(\text{high}[x], \text{max}[\text{left}[x]], \text{max}[\text{right}[x]])$$

In our examples the intervals are closed. An interval tree is ordered by low values of each interval. Thus, the in-order traverse returns the intervals ordered by low borders. It allows insertion, deletion, and search operations. Since an interval tree is a red black tree and height of the red black tree is smaller or equal to $2\lg(n+1)$ (n is the number of nodes (intervals) in the tree), the worst case running time for all three operations is $O(\log(n))$.

Figure 38: An interval tree example. (a) A set of intervals (b) The interval tree constructed from the set of intervals

5.3.1 Interval Tree Operations

- **Search**

A search operation takes an interval such as $[low(x), high(x)]$ and looks for overlapping intervals in the tree.

Fig. 39 shows the pseudo code of interval tree search taken from Cormen et al. [107]. The function takes interval tree T and interval i . It either returns the interval that overlaps to i or the sentinel $nil[T]$. The search starts from the root, continues downward through the tree and terminates when overlapped interval found or x points to the sentinel $nil[T]$.

INTERVAL-SEARCH(T, i)

```
1  $x \leftarrow \text{root}(T)$ 
2 while  $x \neq nil(T)$  and  $i$  does not overlap  $int[x]$ 
3   do if  $left[x] \neq nil[T]$  and  $max[left[x]] \geq low[i]$ 
4     then  $x \leftarrow left[x]$ 
5     else  $x \leftarrow right[x]$ 
6 return  $x$ 
```

Figure 39: Interval tree search pseudo code

- **Insert**

The insert operation takes the root of the interval tree and the interval as an input. It searches for proper position based on the intervals low end and inserts the interval via updating left, right and root pointers of the affected nodes. After insertion, the interval tree fix up function is called to ensure that red-black tree properties are still satisfied.

- **Delete**

The delete operation is similar to insert. It takes the root of the interval tree and an interval as an input. It searches for proper position based on the intervals low end. Along the way the max field of nodes and balance factors are updated.

5.4 Nested Containment Lists

Nested Containment List (NCList), interval database representation, was developed by Alexander V. Alekseyenko et al. [108] to overcome the difficulties of finding overlapping sequence intervals in exponentially growing sequence databases.

Typically time complexity of finding the first overlapping interval to a query is $O(\log N)$ in a database composed of N intervals. But in most cases, interval search requires finding all overlapping intervals. Since intervals are not guaranteed to be contiguous on both start and end coordinates, a search cannot end when the first overlapping interval is found. For example, Fig. 40 demonstrates querying in two databases. In Fig. 42a intervals are ascending ordered based on interval start coordinates. In this type of structure, overlapping intervals might be spread all over the structure. Thus, on average half of the structure needs to be searched to find all overlapping intervals.

If one looks carefully at the start coordinate ascending ordered interval list, it can be seen that intervals contained within another interval prevent them from being sorted based on both start and end coordinates. This problem can be solved by removing contained intervals because if an ascending ordered list of intervals does not have contained intervals, it will be also sorted based on end coordinates. Thus, when the first non-overlapping interval is found, the search can be halted.

Figure 40: Difference between start coordinate sorted database and a NCList data structure. Intervals shown in bold are overlapping intervals with the query interval (figure adapted from [108] under Creative Commons Attribution Non-Commercial License)

Alexander V. Alekseyenko et al. created NCList to partition the intervals into subsists so that in each partition no interval is contained by another interval. For example, in Fig. 42b the first sub-list contains an ascending ordered list of 7 intervals. Additional sub-list intervals that fully overlap with an interval in the predecessor sub-list are shown in a separate box. As seen from the Fig 41b, detected overlapping intervals (shown in black) are located back to back in each of the sub-list. Since the NCList structure guarantees that a sub-list contains only non-overlapping intervals, after a non-overlapping interval to a query interval is found, the search can be halted. In this way the time complexity of finding all overlapping intervals to a query interval is reduced to $O(\log N + n)$. N represents the number of intervals contained in the database of intervals and n represents the number of overlapping intervals to the query interval.

5.4.1 NCList Data Structure

An NCList structure is composed of two arrays; the sub-list headers array (H) and the interval array (L) (Fig 41). The sub-lists headers array (H) stores the record index of

first element and its length for each sub-list S_x . The interval array L stores the sub-lists indexed in H .

Figure 41: “Schematic representation of NCList data structure. Here in the interval array L, element L[i] has non-empty sublist, which records is indexed in sublist header array H by L[i]. SUBLIST. This sublist starts at START and has length LENGTH. L[i] itself belongs to the top level list S0, which has sublist header H[0]” (figure adapted from [108] under Creative Commons Attribution Non-Commercial License)

5.4.2 Query of overlapping intervals

Query of overlapping intervals composed of recursive 3 main steps:

- 1) Binary search in a sub-list S_x .
- 2) Scan in the sub-list to return overlapping intervals that has a sub-list.
- 3) Execute same procedures recursively over the intervals that has a sub-list (Fig 42).

Algorithm 1 *Overlap(L, H, sublist, start, stop)*

```
R ← ∅  
subend ← H[sublist].start + H[sublist].length  
i ← BinarySearchEnd(L, H[sublist], start)  
while i < subend and L[i].start < stop do  
    R ← R ∪ {L[i]}  
    R ← R ∪ Overlap(L, H, L[i].sublist, start, stop)  
    i ← i + 1  
end while  
Return R
```

Figure 42: NCList overlap query pseudo code [108]

The time complexity of this algorithm is $O(\log(N/\alpha) + n)$. N is the size of the intervals database, α is the average number of intervals contained in any top-level interval and n is the number of overlapped intervals for a given query interval.

CHAPTER 6

METHODS

We developed an Affymetrix® GeneChip® probe remapping protocol at the level of genes, transcripts, untranslated regions (UTRs) and individual exons based on the genome and Ensembl annotations. The protocol takes annotations in a General/Gene Transfer Format (GTF) file (see Section 3.3.3), generates a Custom Chip Description File (CDF) where probes are grouped into probe sets based on region (UTR, individual exon), transcript or gene level. Here, we define individual exons as coding exons within protein coding genes, or all exons within structural RNAs (such as miRNA and lncRNA). In effect, the individual exons refer to all non-UTR portions of exons. The region based grouping is beneficial to profile the contribution of UTR, and exon regions to the gene expression levels globally as well as their independent differential expression which may play a significant biological role. The gene and transcript based CDFs allow obtaining gene expression levels of interest based on the latest genomic information. Since annotations rely on the input file provided by the user, it gives researchers the opportunity for analysis based on a specific set of annotations.

6.1 Steps for Generating Custom CDFs

Custom CDF generation is composed of three main steps: mapping probes to the genome, annotation of probes, and assignment of probes to probe sets based on annotations.

Fig. 43 shows the flow chart of annotation and grouping of probes based on the region of a gene. After the probe sequences are mapped to the latest assembled genomic DNA sequences via Bowtie [53] version 1.0.1, the probes were annotated based on the overlapping between probes and gene structures taken from the GTF file using NCList. Probes that share the same annotation were grouped into one probe set. Annotations vary according to the type of CDF file. For region level custom CDFs, probes were mapped to the 5' and 3' UTR, and individual exon start and end locations of every gene and probe sets were named according to the gene id and region. For the gene level custom CDFs, probes were mapped to the start and end locations of every gene and probe sets were named according to gene id. For transcript level custom CDFs, probes were mapped to the start and end locations of every transcript and probe sets named according to transcript id. In every step different filtering criteria are applied for each CDF file.

6.1.1 Mapping of Perfect Match Probes to a Genome

PM probe sequences, which can be obtained from the Affymetrix® Netaffx™ analysis center web site or GEO in a FASTA file format, are aligned to the indexed genome using Bowtie version 1.0.1 with the parameters -v 0 and -m 1, requiring that probes align to a single genomic location with 100% identity, thereby the probes that no longer align to a genomic location or perfectly align to multiple genomic locations are removed and cross-hybridization effects are reduced. Note that Bowtie version 1 is best at aligning shorter sequences (25-50 bp) as found with microarray probes while the most recent versions of Bowtie are optimized for long sequence reads (>50 bp). Mismatch (MM) probes are not considered in the mapping step, although they could theoretically map uniquely to genomic regions. Rather, the MM probes are set aside and are included with

their corresponding PM probe during the final CDF construction step once the PM probes have been assigned to a probe set. During this analysis, only probes perfectly matching to a region are considered. Therefore, probes crossing splice junctions will be discarded. The following is a line taken from the alignment results of Affymetrix® HG-133 Plus 2 to the HG38 genome index.

probe:HG-U133 Plus 2:1053 at:987:1005; Interrogation Position=1216; Antisense; - 7 74232149

Figure 43: Flow chart for region-based probe annotation framework

Until the end of the last semicolon, the line includes the information taken from the Affymetrix® HG-133 Plus 2 probes FASTA file. The rest of the line shows that the probe was aligned to the antisense strand (- for antisense, + for sense strand) on chromosome seven starting at base 74,232,149.

6.1.2 Annotation of Perfect Match Probes via NCList

Probes were annotated based on the overlap between probes and genomic intervals by the following steps.

- I. General/Gene Transfer Format (GTF) files for the studied species were obtained from the Ensembl [86] ftp server. Each GTF is a tab-delimited text file used to represent gene structure information, including the start and end positions of a gene together with chromosome location. Each structure is tagged with a feature which can be gene, transcript, exon, start_codon, stop_codon, CDS or UTR. Fig. 44 shows structure tags marked with red lines for the FO538757.2 gene located in the Homo_sapiens.GRCh38.85.gtf. It has one transcript that composed of three exons. It also has CDS, start_codon, stop_codon and five_prime_utr together with the start and end positions. Ensembl GTFs were used since the annotations are determined by an automated system based on experimentally verified data combined from multiple databases such as RefSeq, EMBL and UniProtKB. It also contains manual curation for selected species.

```

ensembl gene      182393 184158 . + . gene_id "ENSG00000279928"; gene_v
ensembl transcript 182393 184158 . + . gene_id "ENSG00000279928"; ge
ensembl exon      182393 182746 . + . gene_id "ENSG00000279928"; gene_v
ensembl CDS 182709 182746 . + 0 gene_id "ENSG00000279928"; gene_versi
ensembl start codon 182709 182711 . + 0 gene_id "ENSG00000279928"; ge
ensembl exon 183114 183240 . + . gene_id "ENSG00000279928"; gene_v
ensembl CDS 183114 183240 . + 1 gene_id "ENSG00000279928"; gene_versi
ensembl exon      183922 184158 . + . gene_id "ENSG00000279928"; gene_v
ensembl CDS 183922 184155 . + 0 gene_id "ENSG00000279928"; gene_versi
ensembl stop codon 184156 184158 . + 0 gene_id "ENSG00000279928"; ge
ensembl five prime utr 182393 182708 . + . gene_id "ENSG00000279928"

```

Figure 44: Features of the FO538757.2 gene in GTF

- II. A nested containment list (NCList) index was created for each chromosome from intervals (start and end points) of gene structures. The intervals of the NCList were selected based on the target of the probe sets. When the probe sets were constructed based on gene or regions of a gene, we used UTR and individual exon intervals. For transcript targeted probe sets, we used transcript intervals.
- III. Probe intervals were searched in the NCList indexing and annotated according to the overlapping results. Probes were split based on the matched chromosome. Each probe group interval was searched in the same chromosome's NCList indexing. When an overlap was found, the probe was annotated with the list node. Only complete overlaps were accepted; both the low and high ends of the interval must be included in the list node. The probes which did not overlap the nodes were discarded. As a result, probes partially overlapping UTRs, and individual exons will not be included at the region and gene level, but will be present at the transcript level. Table 9 shows the annotations for the Affymetrix® HG-133 Plus 2 probes which were located on chromosome six and searched in the NCList indexing.

TABLE 9: Probe annotation outputs at the end of NCList indexing search

Probe	Start	End	Annotation
223120_at:528:909	143501994	143502018	ENSG000000001036_exon
223120_at:861:491	143501942	143501966	ENSG000000001036_exon
223120_at:683:349	143497416	143497440	ENSG000000001036_exon
223120_at:949:165	143495814	143495838	ENSG000000001036_exon
223120_at:585:237	143495755	143495779	ENSG000000001036_exon
223120_at:605:739	143495514	143495538	ENSG000000001036_exon
223120_at:605:739	143495514	143495538	ENSG000000001036_UTR
223120_at:543:383	143495501	143495525	ENSG000000001036_exon
223120_at:543:383	143495501	143495525	ENSG000000001036_UTR

The following are pseudocode snippets for generating CDFs.

Region Based CDF

```

Get probe intervals and chromosome locations from the text file
Get intervals of individual exons and UTR for each gene from gtf file
Split intervals based on chromosome location
For each chromosome chr {
    Split intervals based on gene region (exon, UTR)
    For each gene {
        For each region
            Combine overlapping intervals
    }
    Create NCList indexing of intervals belong to the chr
    For each probe located on chr
        Search probe interval in the NCList indexing
        Annotate probe with region intersecting gene
    }
Create probe sets
Write probe sets into ASCII CDF files

```

Transcript Based CDF

Get probe intervals and chromosome locations from the text file

Get intervals of transcripts for each gene from gtf file

Split intervals based on chromosome location

For each chromosome chr {

    Create NCList indexing of intervals belong to the chr

    For each probe in chr

        Search probe interval in the NCList indexing

        Annotate probe with intersected transcript

}

Create probe sets

Write probe sets into ASCII CDF files

Gene Based CDF

Get probe intervals and chromosome locations from the text file

Get intervals of individual exons and UTR for each gene from gtf file

Split intervals based on chromosome location

For each chromosome chr {

    Split intervals based on gene region (exon, UTR)

    For each gene {

        For each region

            Combine overlapping intervals

    }

    Create NCList indexing of intervals belong to the chr

    For each probe located on chr

        Search probe interval in the NCList indexing

        Annotate probe with intersected gene of the region

}

Create probe sets

Write probe sets into ASCII CDF files

- IV. A probe's start and end points may overlap multiple genomic structures. It may overlap with the UTR and individual exon region of the same gene (as seen in table 9) or with multiple genes or transcripts. To remove cross hybridization and ensure probes uniquely map to a single region, gene or transcript, we choose one of the annotations for each probe and remove the remaining matches. The rule for assigning these probes occurs with the following priority (I) 5' and 3' UTRs; (II) individual exons. Thus, although UTR regions technically occur within exons, the more specific UTR assignment will be used. When the annotation was based on gene or transcript the first obtained annotation was selected.

In the previous example, probe 223120_at:543:383 overlaps both the exon and UTR region of the *FUCA2* (ENSG00000001036) gene (Table 9). During filtering, matching between the probe and the exon will be removed because our first preference is UTRs (preference order 5' UTR - 3' UTR – individual exon). Because of preference in the CDF the probe will be only used to represent the UTR region of the *FUCA2* gene.

- V. Probes with the same annotation were grouped together to form a probe set.

For example, in the previous example, 223120_at:528:909, 223120_at:861:491, 223120_at:683:349, 223120_at:949:165, 223120_at:585:237 will form a probe set to represent the individual exon region of the *FUCA2* gene. Fig. 45 shows the grouping of probes for three types of CDFs. These CDFs are:

- Region-based CDF: Probe sets are designed to target a specific region of a gene and consist of probes which map to the same region (UTR, individual exon) of

a gene. In Fig. 45, green probes were mapped to the UTR region of Gene_1; therefore, those probes cluster together to form the Gene_1 UTR region probe set. Based on the same logic, blue colored probes form the probe set for Gene_1 exon.

- Gene-based CDF: Probe sets are designed to target genes and consist of probes which map to the same gene. In Fig. 45, green and blue colored probes, which mapped to Gene_1, cluster together to form the Gene_1 probe set.
- Transcript-based CDF: Probes that map to same transcript of a gene compose a probe set. In Fig. 45, the orange and red arrow show the start and end positions of Transcript_1 and Transcript_2. The probes mapped to the Transcript_1 (two greens, five blues) cluster together to form the probe set for Transcript_1.

Figure 45: Creating probe sets for different types of custom CDF based on probe mapping to genomic structures

VI. Probe sets were saved into ASCII format CDF files. In addition to the probes specific for a particular gene, Affymetrix® GeneChips® contain a number of different control

probes such as probes that are added during sample preparation, providing evidence that assay was performed properly. We added those probe sets to our CDFs without any change.

6.1.3 Probe Set Naming

Since GTF files obtained from Ensembl were used, Ensembl gene ids were employed to distinguish different genes and Ensembl transcript ids were used to distinguish different transcripts. When the generated CDF was based on regions of genes, the region was suffixed to the Ensembl gene id. Strand direction specified in the GTF of the genomic structure is also suffixed (- or +). Table 10 shows example probe set names taken from custom CDFs for the Affymetrix® GeneChip® HG-133 Plus 2.

TABLE 10: Custom CDF naming examples

CDF Type	Probe Set Name
Region-based	ENSG00000001036_EXON_-
	ENSG00000001631_THREE_PRIME_UTR_-
Gene-based	ENSG00000001461_+
Transcript-based	ENST00000489806_+

CHAPTER 7

affyCustomCdf: AN R PACKAGE FOR CREATING CUSTOM CDF

We developed an R package, `affyCustomCdf`, for allowing users to create custom CDFs via removing nonspecific probes, updating probe target mapping based on the supplied genome information and grouping probes into region (individual exon, UTR), gene and transcript levels. It has been implemented in R and C++ languages based on the Bioconductor package requirements and publicly available at <https://github.com/UofLBioinformatics/affyCustomCDF>.

7.1 `affyCustomCdf` Tool

To create a custom CDF via `affyCustomCdf` tool, one must call the `createAffyCustomCdf` function after installation of the tool. The `createAffyCustomCdf` function produces a custom CDF based on parameters this includes the original CDF, probe alignment file and annotation (GTF) file which must be supplied by the user. The remaining parameters are optional, with a default value set (for more details see appendix). Fig. 46 describes the parameter selection for creating a custom CDF via the `affyCustomCdf` tool.

7.1.1 Custom CDF Types

The `affyCustomCdf` tool can produce main three types of custom CDF. They are region-based, gene-based and transcript based.

- **Region based CDF:** In region based custom CDFs, probe sets consist of probes which map to the same region (individual exon, UTR) of a gene. It can be obtained via supplying “regionG” value to the type parameter of the createAffyCustomCdf function.
- **Gene based CDF:** In gene based custom CDFs, probe sets consist of probes which map to the same gene. It can be obtained via supplying “gene” value to the type parameter of the createAffyCustomCdf function.
- **Transcript based CDF:** In transcript based custom CDFs, probe sets consist of probes which map to the same transcript of a gene. It can be obtained via supplying “transcript” value to the type parameter of the createAffyCustomCdf function.

In addition to main three types supplied above, files can be diversified via defining rules for probe selection and directional relationship between probes and annotations.

Three different types of custom CDF can be produced based on the sense/antisense relationship between probes and annotations. They are no direction, same direction and opposite direction.

- **No direction CDF:** When no direction rule is applied, direction will not be considered during probes to annotations mapping. This can be performed via supplying “0” value to the SD parameter of the createAffyCustomCdf function.
- **Same direction CDF:** When same direction rule is applied, mapping is performed between same directions such as sense probes can only map to sense annotations. This can be performed via supplying “1” value to the SD parameter of the createAffyCustomCdf function.

- **Opposite direction CDF:** When opposite direction rule is applied, mapping is performed between opposite directions such as sense probes can only map to antisense annotations. This can be performed via supplying a value except 0 or 1 to the SD parameter of the createAffyCustomCdf function.

Probe selection can be defined in two ways. Allowing probe share between transcripts of a gene or removing a probe according to the number of matched annotations to the probe.

- **Use probes map to one annotation:** When unique probe option is selected, probes that maps to more than one annotation are discarded from a region-based CDF. This option is provided to allow users to examine unique probes to one specific annotation. It can only be used with the regionG type. Selection can be supplied to the createAffyCustomCdf function via uniqueProbe parameter. Possible values are FALSE for not unique probes, TRUE for unique probes.
- **Share probe between transcripts of a gene:** In the case of transcript based custom CDFs, when an exon is shared between transcripts of a gene, only one of them can have probes that maps to the exon because we choose one of the annotations that a probe maps and remove the remaining matches (See chapter 6 for more details). Via share probe between transcript of a gene option, probe sharing between transcripts of a gene is allowed. This option is provided to cover more transcripts in the CDF and supply more accurate predictions for every transcript of a gene via having more probes per probe set.

The last option that creates diverse CDF is junction tolerance.

- **Junction tolerance file:** Junction tolerance allows CDFs to include probes that map to the junctions of annotations via merging overlapping and consecutive intervals of same structures before creating NCList indexes. For example, in Rat assembly rn6 the *PRR4* gene has four transcripts. One of the transcripts (ENSRNOT00000078821) has an exon (ENSRNOE00000523383) that starts at 167,483,620 and ends at 167,483,880. Another transcript (ENSRNOT00000038396) has an exon (ENSRNOE00000557210) that starts at 167,483,881 and ends at 167,483,910. When junction is not allowed, a probe in Rat 230 2.0 microarray with 167,483,620 and 167,483,887 coordinates will not be annotated with PRR4-exon because it does not align to an individual exon. When junction is allowed a new interval 167,483,620 and 167,483,910 will represent the PRR4-exons and the probe will be annotated and placed in the produced custom CDF.

7.1.2 Interpreting Results

The affyCustomCdf tool produces three files. They are custom CDF file, report file and missing probes file.

- **Custom CDF file:** Produced custom CDF file consists of regrouped probes into probe sets based on specific annotations of interest. It is used for data analyses of Affymetrix® GeneChip® data.
- **Report File:** Report file consists of multiple information. They are:
  - Parameter values sent to the CreateAffyCustomCdf function.
  - Distribution of number of probes per probe set table.
  - The histogram of the number of probes per probe set.

- Number of probes aligned to an annotation.
- Number of annotation detected by probes before probe sharing eliminated.
- **Missing Probes File:** In some cases, the FASTA file of probes may include some probes that is not used in the original CDF. These probes are detected and placed into a text file to inform users. The name of the file starts with the original CDF name with postfix missingProbes. i.e Rat230_2.cdf_missingProbes.txt

7.1.3 Running Time

As long as the version of a genome and annotations are not changed, probe sets designed inside the custom CDFs are not required to be created more than one time providing that in Affymetrix® GeneChips® probes do not change. Therefore, long running time does not become an important issue. But still we know that short running time is always a positive feature to have in a tool. To achieve relatively fast tool, we took the advantage of NCList indexing and hash tables. Moreover, we implemented some part of the tool with C++ and combined these parts with R code. Table 11 shows the running time result in seconds performed on a computer with eight GB memory and 2.80 GHz CPU.

TABLE 11: Running times

	Region Based	Gene Based	Transcript Based
Human U133 Plus 2.0 Array	23 min	23 min	4 min
Rat Genome 230 2.0 Array	6 min	6 min	71.82 sec
Mouse Genome 430 2.0 Array	13 min	14 min	3 min

Figure 46: affyCustomCdf parameter selection

CHAPTER 8

EXPERIMENTAL RESULTS

In this section we demonstrate the remapping and probe set results for some of our custom CDFs. We also include analysis results performed by using our custom CDFs. The experiments were performed on the Affymetrix® GeneChip® HG-133 Plus 2, Rat 230 2.0. and Mouse 430 2.0. During the creation of custom CDFs, no direction rule is applied and junction is not allowed (see Section 7.1.1 custom CDF types).

8.1 Probes Mapping to Genome

Using the bowtie version 1.0.1 we were able to identify probes that uniquely map with 100% identity for each of the respective genomes. As a result, 87% PM probes of the HG-U133 Plus 2, 84% PM probes of the Rat 230 2.0 and 86% PM probes of the Mouse 430 2.0 were uniquely mapped to the genome and were used in the subsequent steps (Table 12).

TABLE 12: Number of mapped probes for custom CDF construction

GeneChip®	# of PM Probes	# of PM Probes Mapped Uniquely	# of PM Probes Mapped to Multiple Locations	# of PM Probes Not Aligned
Human Genome U133 Plus 2.0 Array	603,158	525,985	364,93	406,80
Rat Genome 230 2.0 Array	341,459	288,319	260,27	271,13
Mouse Genome 430 2.0 Array	495,374	427,758	284,44	391,73

8.1.1 HG-133 Plus 2

Chromosome sequences of the human assembly 38 (October 2014) were obtained from the Ensembl ftp and bowtie indexes were created using bowtie-build version 1.0.1 with default parameters. We also obtained the DNA sequences for perfect match (PM) probes of Affymetrix® GeneChip® HG-U133 Plus 2 from the Affymetrix® Netaffx™ web site as a FASTA file format. After we removed the control probes from FASTA file, we performed mapping of PM probes to the indexed genome using Bowtie version 1.0.1 (parameters $-v\ 0 - m\ 1$). We filtered all the probes except those that aligned to one genomic location with 100% identity. As a result, 525,985 out of 603,158 PM probes uniquely mapped to the genome and were used in the next steps. 36,493 probes mapped to multiple locations and were removed. 40,680 PM probes did not align to the genome without mismatches and were also removed.

8.1.2 Rat 230 2.0

Chromosome sequences of rat assembly Rnor 6 were obtained from NCBI and created bowtie indexes using bowtie-build version 1.0.1 with default parameters. We also obtained the DNA sequences for the perfect match (PM) probes of Affymetrix® GeneChip® Rat 230 2.0 from the Affymetrix® Netaffx™ web site in FASTA file format. After we removed the control probes from the FASTA file, we performed mapping of PM probes to the indexed genome using Bowtie version 1.0.1 (parameters $-v\ 0 - m\ 1$). We filtered out all the probes except ones that aligned to one genomic location with 100% identity. As a result, 288,319 out of 341,459 PM probes uniquely mapped to the genome and were used in the next steps. 26,027 probes mapped to multiple locations and were

removed. 27,113 PM probes did not align to the genome without mismatches and also were removed.

8.1.3 Mouse 430 2.0

Chromosome sequences of Mouse assembly mm10 were obtained from the Ensembl ftp and created bowtie indexes using bowtie-build version 1.0.1 with default parameters. We also obtained the DNA sequences for the perfect match (PM) probes of Affymetrix® GeneChip® Mouse 430 2.0 from the Affymetrix® NetAffx™ web site in FASTA file format. After we removed the control probes from the FASTA file, we performed mapping of PM probes to the indexed genome using Bowtie version 1.0.1 (parameters $-v\ 0 - m\ 1$). We filtered out all the probes except ones that aligned to one genomic location with 100% identity. As a result, 427,758 out of 495,374 PM probes uniquely mapped to the genome and were used in the next steps. 28,444 probes mapped to multiple locations and were removed. 39,173 PM probes did not align to the genome without mismatches and also were removed.

8.2 Probe Annotations and Probe Sets

To annotate probes, we mapped uniquely aligned probes to gene regions using the Ensembl GTF file for each respective organism. We selected the specific regions from GTF based on the custom CDF type (gene, transcript or region-based). Consequently, we produced three types of custom CDFs (Tables 13 and 14).

TABLE 13: Summary of probes used for custom gene and transcript based Custom CDFs

	<i>Homo sapiens</i>		<i>Rattus norvegicus</i>		<i>Mus musculus</i>	
	Gene	Transcript	Gene	Transcript	Gene	Transcript
# of Probes Used	414,701	504,419	162,356	205,671	323,917	395,884
# of Probe Sets Constructed	226,51	260,96	131,50	144,66	192,82	209,80
Average # of Probes Per Probe Set	18	18	12	14	16	18

TABLE 14: Summary of probes used for region based custom CDFs

	<i>Homo sapiens</i>	<i>Rattus norvegicus</i>	<i>Mus musculus</i>
# of Probes Aligned to Genome	822,681	321,905	637,942
# of Probes Used	414,701	162,356	323,917
# of Probe Sets Constructed	339,16	198,39	289,63
Average # of Probes Per Probe Set	12	8	11

8.2.1 Detailed Results for HG-133 Plus 2

We used the Ensembl GRCh38 genome version's GTF file for the annotations and created three different CDF files for future data analysis.

I. Region Based Custom CDF

After mapping probes to the genomic regions, we identified 822,681 annotations (a probe aligned to a genomic region and marked with this region). The custom CDF has 33,939 custom designed probe sets: 16,375 probe sets represent UTR regions and 16,024 probe sets represents exon, plus 62 original control probe sets. 405,020 annotations were filtered.

Fig. 47 shows the distribution of probe numbers per probe set. The minimum probe number per probe set is 1, the average number of probes per probe set is around 12 and the median is 11.

Pie Chart of Probe Number Per Probe Set Gene Region CDF

Figure 47: The distributions of probe numbers per probe set for the gene region based CDF

II. Gene Based Custom CDF

The human gene based CDF has 22,651 custom designed probe sets composed from 414,692 probes and 62 original control probe sets. 442,025 annotations were identified between genes and the probes. 27,324 annotations were filtered after shared probes were removed.

In order to validate our probe set annotations, we compared the original CDF probe sets with the custom CDF. A total of 21,585 annotated genes were shared between the two CDFs, with 3,068 unique to the original CDF, and 1,066 unique to our custom CDF. We obtained the Ensembl id for each probe set from the Ensembl Funcgen database [109] using Perl scripts. Fig. 48 shows the comparison results. 85% (except control probes) of the original probe targets are preserved in the custom CDF.

Figure 48: Original CDF versus custom CDF

To determine why some genes were not covered in our CDF, we examined those unique to the original CDF. First, we obtained the probe sets which represent these genes in the original CDF, yielding 2,781 probe sets. We retrieved both the PM and MM probe

sequences for each of these. We observed that for 667 probe sets, every probe was removed during probe mapping to the genome due to either non-unique mappings or mapping rates less than 100%. 30,150 probes from the remaining 2,114 probe sets were not used in our CDF since they either did not map to the genome or they were MM probes. 14,028 probes were used in our newly constructed probe sets which target different genes than the original assignment by Affymetrix® and 2,656 probes were not aligned to gene structures and not annotated. As a result, the differences between the original CDF and our method occurs because of probes removed during genome alignment, probes that no longer map to gene structure or probes that map to gene structures different from the original annotation.

Fig. 49 shows the distribution of probe numbers per probe set. The minimum probe number per probe set is 1, the average number of probes per probe set is around 18 and the median is 11.

Pie Chart of Probe Number Per Probe Set Gene CDF

Figure 49: The distributions of probe numbers per probe set for the gene based CDF

We also compared our gene based CDF with the Brain Array Ensembl Gene version 19 CDF (Fig. 50). Most of the genes that were not included in our CDF were removed during the filtering of multiple annotated probes. A few of the missing genes are located on the mitochondrial Chromosome MT which is not included in our Bowtie index. We annotated an additional 2,758 genes.

Figure 50: Original CDF versus the brain array CDF

III. Transcript Based Custom CDF

The unique probes transcript based CDF has 26,096 custom designed probe sets composed from 504,419 probes. We also included 62 original control probe sets. 2,129,643 annotations were identified between transcripts and the probes.

Fig. 51 shows the distribution of probe numbers per probe set. The minimum probe number per probe set is one, the average number of probes per probe set is around nineteen and the median is eleven.

Figure 51: The distributions of probe numbers per probe set for the transcript based CDF

8.3 Analysis Results

To test our custom CDFs, we reanalyzed the publicly available data series GSE72551 and GSE48611. Both studies involve the nervous system, where differences in 3' UTRs are likely to have phenotypic effects on transcript localization. We tested it on the publicly available data series. The GSE48611 data set was analyzed via HG-133 Plus 2 custom CDFs. The GSE48611 was analyzed via Rat 230 2.0 custom CDFs. Prior to analysis, we removed probe sets with two or fewer probes from the custom CDFs in order to achieve more accurate results for target expression levels. Robust Multiarray Averaging (RMA) normalization was used for preprocessing. A p-value 0.05 was used as the threshold for all experiments.

- **Results for GSE48611**

The GSE48611 [63] was obtained from GEO. The data series examines Down Syndrome (trisomy 21) gene expression monitoring. They have collected mRNA samples from the isogenic trisomy of chromosome 21 (Ts21) and control pluripotent stem cells (iPSCs) (DS1, DS4, and DS2U) between passages 24 and 48 and from day 30 neurons. They have three biological replicates for each case. In our analysis we looked for the differentially expressed genes (DEGs) between isogenic Ts21 and control iPSCs for DS1 and DS4. To identify the DEGs, we set an adjusted p-value threshold of 0.05 for all experiments and obtained the genes passing this threshold. Before starting the analysis, we removed the probe sets which have 3 or fewer probes per probe set from custom CDFs. First, we analyzed the data based on our gene based custom CDF and the original Affymetrix® supplied CDF obtained from the Affymetrix® Netaffx™ web site. For DS1, our gene based CDF identified 814 DEGs, while the original CDF identified 2,402 DEGs

with 625 DEGs identified by both methods (Fig. 52). Table 15 shows the top ten DEGs and their ranking in the original method results. Since the original CDF has multiple probe sets per gene, we only reported the top ranking for each gene and adjust the rankings. We have two genes (*ZNF717*, *ZNF560*) in our top ten list from qPCR verified genes which is supplied in the original paper.

We repeated the same analysis for DS4 versus DS2U. Our gene base CDF identified 621 DEGs, while the original CDF identified 1,819 unique DEGs with 379 DEGs identified by both methods (Fig. 53). Table 16 shows the top ten DEGs based on our CDF and their adjusted ranking in the original method results. We have three genes (*ZNF717*, *ZNF560*, *CRYZ*) in our top ten list from qPCR verified genes which is supplied in the original paper. Eight genes were commonly expressed for DS1 and DS4.

Figure 52: Typical method versus our method for DS1 versus DS2U

TABLE 15: DS1 versus DS2U top ten DEGs

Our Method				Original Method	
Gene Id	logFC	adj.P.Val	Ranking	logFC	adj.P.Val
<i>ZNF667-AS1</i>	5.682693	1.50E-08	1	5.617611	5.53E-09
<i>CTSF</i>	1.906976	4.24E-08	2	2.074571	2.89E-08
<i>ZNF667</i>	3.579938	4.19E-07	4	3.910541	8.74E-07
<i>ZNF717</i>	3.209391	6.06E-07	16	2.355922	2.96E-05
<i>NNAT</i>	3.445	2.62E-06	18	2.125069	3.45E-05
<i>ZNF560</i>	2.764886	5.54E-06	7	2.638967	1.97E-06
<i>TMSB15A</i>	-2.28818	6.54E-06	NA	NA	NA
<i>ZNF578</i>	1.990097	1.80E-05	19	1.940997	3.52E-05
<i>ZNF239</i>	-2.1717	1.80E-05	13	-2.19910	2.39E-05
<i>PAXBP1</i>	1.057482	1.86E-05	23	1.000226	4.57E-05

Figure 53: Typical method versus our method for DS4 versus DS2U

TABLE 16: DS4 versus DS2U top ten DEGs

Our Method				Original Method	
Gene Id	logFC	adj.P.Val	Ranking	logFC	adj.P.Val
<i>ZNF717</i>	3.316176	1.50E-06	8	2.564958	2.25E-05
<i>ZNF667-AS1</i>	3.97883	1.57E-06	1	3.996053	1.14E-06
<i>NNAT</i>	3.591202	2.38E-06	10	2.221193	4.02E-05
<i>ZNF578</i>	2.205069	8.71E-06	7	2.167641	2.25E-05
<i>CTSF</i>	1.233071	1.04E-05	6	1.292598	1.34E-05
<i>NLRP2</i>	1.575696	1.04E-05	5	3.454983	8.77E-06
<i>PAXBP1</i>	0.977935	7.41E-05	19	0.862839	0.000480611
<i>CRYZ</i>	5.230353	7.41E-05	11	5.240513	7.67E-05
<i>ZNF560</i>	2.220455	8.07E-05	9	2.166272	3.74E-05
<i>SRP19</i>	-0.86613	0.000161	33	-0.81259	0.001628958

We analyzed DS1 versus DS2U via our region based CDF as well. The CDF identified 1,494 DE gene regions. 56 DE gene regions target the same gene but different regions (UTR and exon) (Fig. 54). 698 of them only come from UTR and 740 of them did not have probe sets in UTR region (i.e. they only come from exon regions). Fig. 55 shows the result for DS4 versus DS2U.

Figure 54: Distributions of probe set regions in the DS1 versus DS2U DEGs

Figure 55: Distributions of probe set regions in the DS4 versus DS2U DEGs

In order to see how the region based CDF changes the analyses, we took the Ensembl id of every region probe set and compared them with the Ensembl id of the gene based result (Fig. 56, Fig 57). It showed that the number of commonly detected DEGs is small compared to the number of detected DEGs in each case.

Figure 56: Distribution of common and different DEGs for our gene and gene region based methods in the DS1 versus DS2U case

Figure 57: Distribution of common and different DEGs for our gene and gene region based methods in the DS4 versus DS2U case

Table 17 shows the top ten DEGs for DS1 versus DS2U. Table 18 shows the top ten DEGs for DS4 versus DS2U. In the case of Ds4 versus the DS2U region based three genes of top ten are same with D1 versus DS2U region based.

TABLE 17: DS1 versus DS2U top ten region based DEGs

		Region Method		Gene Method	
Probe Name	Gene Id	logFC	adj. P.Val	Ranking	adj.P.Val
ENSG00000154646_UTR_-	<i>TMPRSS15</i>	1.618164	4.33E-06	13712	0.763167
ENSG00000143473_exon_-	<i>KCNH1</i>	1.506579	4.33E-06	15435	0.825465
ENSG00000135916_UTR_+	<i>ITM2C</i>	3.143009	1.77E-05	7385	0.468028
ENSG00000232837_exon_-	<i>AF064858.7</i>	1.312395	0.000157	12364	0.709271
ENSG00000122085_exon_-	<i>MTERFD2</i>	1.582542	0.000157	6532	0.419254
ENSG00000229913_exon_-	<i>RP11-378I13.1</i>	1.778681	0.000157	13950	0.771693
ENSG00000184361_UTR_-	<i>SPATA32</i>	0.896036	0.000157	21500	0.998295
ENSG00000224418_exon_+	<i>STK24-AS1</i>	0.912373	0.000192	12250	0.704311
ENSG00000174915_UTR_+	<i>PTDSS2</i>	1.644557	0.000215	10517	0.630927
ENSG00000204179_UTR_+	<i>PTPN20</i>	1.340051	0.000215	3403	0.212167

TABLE 18: DS4 versus DS2U top ten region based DEGs

Probe Name	Gene Id	Region Method		Gene Method	
		logFC	adj P.Val	Ranking	adj. P.Val
ENSG00000154646_UTR_-	<i>TMPRSS15</i>	1.117224	0.000477	859	0.058301
ENSG00000215644_exon_+	<i>GCGR</i>	5.114162	0.000477	18359	0.921109
ENSG00000143473_exon_-	<i>KCNH1</i>	1.044703	0.000477	14504	0.790649
ENSG00000265491_exon_-	<i>RNF115</i>	-1.2177	0.000557	16520	0.862608
ENSG00000224418_exon_+	<i>STK24-AS1</i>	0.857835	0.000603	15478	0.828281
ENSG00000160294_exon_-	<i>MCM3AP</i>	1.042106	0.000603	123	0.010939
ENSG00000113384_UTR_-	<i>GOLPH3</i>	1.170004	0.000737	18153	0.799779
ENSG00000132016_UTR_-	<i>C19orf57</i>	1.451636	0.000831	8375	0.505504
ENSG00000099219_UTR_-	<i>ERMP1</i>	-0.98893	0.000869	16003	0.849883
ENSG00000227234_exon_+	<i>SPANXB1</i>	-1.01635	0.000869	19103	0.941769

We also used the Brain Array Ensembl Gene version 19 custom CDF to compare our method's gene based results. 712 genes detected as DE by both methods (Fig. 58). Our method detected 189 unique DEGs and Brain array detected 182 unique DEGs.

Figure 58: DEGs detected by the Brain array method and our gene based method for DS1 versus DS2U

TABLE 19: D1 versus DS2U top ten upregulated genes Brain Array Ensembl V 19

External Id	logFC	adj.P.Val	Our Method Rank
<i>CTSF</i>	1.897594	1.06E-07	2
<i>ZNF667</i>	3.530854	6.31E-07	3
<i>ZNF667-AS1</i>	5.134179	1.22E-06	1
<i>NNAT</i>	2.84469	2.68E-06	5
<i>ZNF560</i>	2.741947	6.39E-06	6
<i>TMSB15A</i>	-2.29051	6.89E-06	7
<i>ZNF239</i>	-2.27282	6.89E-06	9
<i>PAXBP1</i>	1.129587	1.52E-05	10
<i>ZNF578</i>	1.986575	1.68E-05	8
<i>UTY</i>	1.968354	2.81E-05	11

Nine out of ten genes correlate with our results. The UTY gene is detected as the 11th DEG.

- **Results for GSE72551**

The GSE72551 data series examines gene expression changes associated with collateral sprouting and includes 5 naive controls, 7 replicates at day 7 post-surgery and 7 replicates at day 14 post-surgery. In our analysis we looked for the differentially expressed genes (DEGs) between 0-7 and 0-14. To identify the DEGs, we set a p-value 0.05 as the threshold for all experiments and obtained the genes passing this threshold. Before starting the analyses, we removed the probe sets which have less than three probes from custom CDFs. First, we analyzed the data based on our gene based custom CDF and original Affymetrix® supplied CDF obtained from the Affymetrix® Netaffx™ web site. For 7 versus 0, our gene base CDF identified 981 DEGs; original CDF identified 1,695 unique DEGs. 750 DEGs identified by both methods (Fig. 59). Table 20 shows the top ten DEGs and their ranking in the original method results. For 14 versus 0, our gene base CDF identified 1,948 unique DEGs, original CDF identified 3,484 unique DEGs. 1,560 DEGs

identified by both methods (Fig. 60). Table 21 shows the top ten DEGs and their ranking in the original method results.

Figure 59: Typical method versus our method for day 7 versus naive

Figure 60: Typical method versus our method for day 14 versus naive

TABLE 20: Day 7 versus naive top ten DEGs

Our Method			Original Method		
Gene Id	logFC	P.Val	Ranking	logFC	P.Val
<i>Slc41a2</i>	0.327865	1.87E-05	5	0.331196	4.05E-05
<i>Txnrd1</i>	0.30567	2.16E-05	13	0.323269	7.61E-05
<i>Nppa</i>	0.400947	4.26E-05	8	0.410406	5.19E-05
<i>Grhl3</i>	-1.11758	4.97E-05	3	-1.02727	2.80E-05
<i>Bdnf</i>	0.754445	5.93E-05	10	0.793467	6.86E-05
<i>Rasgrp2</i>	-0.34898	6.34E-05	45	-0.38074	0.000517
<i>Abca8a</i>	-0.29425	0.000104	15	-0.28487	0.000117
<i>Pdzd3</i>	-0.37835	0.000114	11	-0.38425	6.98E-05
<i>Gabra5</i>	0.721666	0.000138	15	0.742534	8.59E-05
<i>Cat</i>	-0.29616	0.000156	22	-0.29166	0.00021

TABLE 21: Day 14 versus naive top ten DEGs

Our Method			Original Method		
Gene Id	logFC	P.Val	Ranking	logFC	P.Val
<i>AABR07026969.1</i>	-0.56774	8.43E-07	5	0.331196	4.05E-05
<i>Bbs5</i>	-0.54004	6.47E-06	27	-0.4116	2.94E-05
<i>Pde7b</i>	0.374652	7.25E-06	29	0.339101	3.49E-05
<i>Epb41l4a</i>	-0.37437	1.58E-05	16	-0.3719	1.41E-05
<i>Tgs1</i>	-0.39079	1.95E-05	35	-0.38539	5.32E-05
<i>Txnrd1</i>	0.307737	1.99E-05	64	0.308068	0.000131
<i>Pcsk6</i>	-0.28696	2.37E-05	24	-0.28712	2.77E-05
<i>Wdr75</i>	-0.3425	2.40E-05	34	-0.38832	5.17E-05
<i>Tmem18</i>	-0.37817	3.31E-05	2018	-0.23385	0.009632
<i>Cish</i>	0.293301	4.15E-05	162	0.41206	0.000387

We analyzed the data via region based CDF as well. For condition 7, the CDF identified 1,350 DE gene regions. 94 DE gene regions target the same gene but different regions (UTR and exon) (Fig. 61). 695 of them only come from UTR and 561 of them did not have probe sets in UTR region (i.e. they only come from exon regions).

Figure 61: Distributions of probe set regions in the 7 versus naive DEGs

For condition 14, the CDF identified 2,573 DE gene regions. 248 DE gene regions target the same gene but different regions (UTR and exon) (Fig. 62). 1349 of them only come from UTR and 976 of them did not have probe sets in UTR region (i.e. they only come from exon regions).

Figure 62: Distributions of probe set regions in the 14 versus naive DEGs

In order to see how the region based CDF changes the analyses, we took the Ensembl id of every region probe set and compared them with the Ensembl id of the gene

based result (Fig. 63, Fig. 64). It showed that the number of common detected DEG is small compare to number of detected DEGs in each case.

Figure 63: Distribution of common and different DEGs for our gene and gene region based methods in the 7 versus naive case

Figure 64: Distribution of common and different DEGs for our gene and gene region based methods in the 7 versus naive case

Table 22 shows the top ten DEGs for 7 versus naive. Table 23 shows the top ten DEGs for 14 versus naive.

TABLE 22: 7 versus naive top ten region based DEGs

Probe Name	Gene Id	Region Method		Gene Method	
		logFC	P.Val	Ranking	P.Val
ENSRNOG00000018225_UTR_+	<i>Tp53inp2</i>	0.445475	2.34E-05	7909	0.573936
ENSRNOG00000002327_UTR_-	<i>Gabrb1</i>	0.813037	2.69E-05	10404	0.805792
ENSRNOG000000045989_exon_-	<i>LOC287167</i>	0.628027	3.07E-05	8949	0.666452
ENSRNOG000000020076_exon_+	<i>RGD1311783</i>	0.32314	4.15E-05	738	0.03043
ENSRNOG00000002336_exon_+	<i>Gabra4</i>	0.736634	4.47E-05	2792	0.155193
ENSRNOG00000016847_UTR_+	<i>Bace1</i>	0.756573	5.99E-05	6593	0.46302
ENSRNOG00000014312_exon_+	<i>TRBC2</i>	-0.87884	7.60E-05	8180	0.596293
ENSRNOG000000060599_UTR_+	<i>Gabrb3</i>	0.645368	0.000113	1281	0.058284
ENSRNOG00000006027_exon_+	<i>Eif2ak4</i>	0.308961	0.000129	6730	0.476479
ENSRNOG00000010361_UTR_+	<i>Kif3b</i>	0.23965	0.000148	4056	0.251742

TABLE 23: 14 versus naive top ten region based DEGs

Probe Name	Gene Id	Region Method		Gene Method	
		logFC	P.Val	Ranking	P.Val
ENSRNOG00000020877_UTR_-	<i>Pyy</i>	-0.77755	5.943563	521	0.005948
ENSRNOG00000024452_UTR_+	<i>Rltpr</i>	-0.60557	6.594111	9979	0.681486
ENSRNOG000000054080_exon_-	<i>Cgnl1</i>	-0.45831	8.026746	11422	0.861384
ENSRNOG000000060931_exon_-	<i>Rspy1</i>	-0.36454	10.27095	1096	0.017312
ENSRNOG00000002327_UTR_-	<i>Gabrb1</i>	0.839938	8.305223	2309	0.051623
ENSRNOG00000015231_UTR_+	<i>Mrpl44</i>	0.284317	9.885974	7862	0.446372
ENSRNOG00000025496_exon_+	<i>Usp7</i>	-0.72064	5.742713	8428	0.501202
ENSRNOG00000002336_exon_+	<i>Gabra4</i>	0.773239	8.259599	301	0.00283
ENSRNOG00000013956_exon_-	<i>Rnf38</i>	-1.34435	6.044207	9180	0.586421
ENSRNOG00000014312_exon_+	<i>TRBC2</i>	-0.94034	5.994768	2577	0.062007

We also reanalyzed the data using brainarray Ensembl CDF version 20. Fig. 65 shows a Venn diagram representing the number of differentially expressed genes using region, gene and brainarray custom CDFs for both cases.

Further examination of the 7 day versus naive ENSEMBL genes found to be differentially expressed in either the gene-based or region-based CDF shows high

concordance, with 975 ENSEMBL genes determined to be differentially expressed using both CDFs (Fig. 65a). Examination of the p-values shows a significant correlation between both the gene and the 3' UTR region ($r=0.439$; $p=1.480E-58$) as well as the gene and the exon region ($r=0.101$; $p=0.001$). The higher correlation with the 3' UTR region is to be expected, due to a higher abundance of probes designed in these regions.

160 genes are found to be differentially expressed using the gene-based approach only, with three not included in the region-based CDF. Further examination shows that 122 of these (78%) have a gene-based pvalue > 0.03 , and 80 (50%) have a gene-based p-value > 0.04 , indicating the detected differences are just below the cutoff level. Analysis of the region-based p-values show that 120 of these (77%) have a region based p-value < 0.10 , and 146 (94%) have a region-based p-value < 0.20 , putting these genes just above the significance threshold.

An additional 423 genes are found to be differentially expressed using the region-based approach only, with 203 from the 3' UTR only, 10 from the 5' UTR only, 206 from the exon only, and 4 from both the 3' UTR and exon. Unlike the DEGs uniquely found in the gene-based approach, those genes found to be differentially expressed in the region-based approach typically have a much higher p-value in gene-based analysis, with only 31% having a p-value between 0.05 and 0.10. This supports our reasoning that separating into functional regions allows detection of subtle changes in transcript formation that may have a larger functional impact of those transcripts which has been further validated by experimental work showing differential expression of the 3' UTR of the CAMKIV gene plays a role in localization [110].

(a) 7 versus naïve

(b) 14 versus naïve

Figure 65: Number of common and different differentially expressed genes using region, gene and brain array custom CDFs. a Day 7 versus naïve. b Day 14 versus naïve

In order to determine why some genes were only detected by brainarray, we examined probes of those genes. 39 probes were not used in our CDFs since they aligned to multiple locations in the rn6 genome. 10 probes did not match gene structures in Ensembl

and were not used in the CDFs. 18 probes were removed because the probe set contained fewer than three probes. 40 probes were used in different probe sets other than those annotated by brainarray.

8.4 Conclusions

One of the limitations of microarray technologies is the design of probes based on available sequence and annotation data at the time of design. Based on our analysis, the percentage of uniquely mapping probes varies from 84% (rat) to 87% (human), indicating that changing knowledge about the genome itself plays a role in probe utilization. In terms of annotation, the rat genome is known to have more incomplete information when compared to mouse and human, which is reflected in the fact that only 47% of the rat probes lie in region-based locales (exons and UTRs) compared to 65% for mouse, and 69% for human. Since this can potentially lead to a small number of probes in each annotated region (and thus increased false positive rates), we have further required at least three probes be present in each probe set for our analysis. To further illustrate the importance of region-based CDFs, using the subset of 4,024 genes with probesets in both the individual exon and 3' UTR regions, we were able to identify 203 differential expression events at the 3' UTR level that do not show differential expression within the exon. In addition, these events are not detected using the standard Affymetrix® CDF. Further analysis of these 203 genes yields some genes of particular interest. For instance, the 3' UTR of *GRIK4* (Glutamate Ionotropic Receptor Kainate Type Subunit 4) was up-regulated (p-value 0.0450) while the CDS was not significantly regulated (FC=1.07; p-value 0.4525), suggesting the 3' UTR of this gene was lengthened (Fig. 66). *GRIK4* regulates kainite-receptor signaling and neuroplasticity [111] and its missregulation is associated with

neurological diseases including Alzheimer's, bipolar disorder [112], and others. Interestingly, a deletion variant specific to the 3' UTR of *GRIK4* is protective of bipolar disorder[112]. Alongside our observation, this shows an example of how a gene can be regulated through the 3' UTR. We also observed that the 3' UTR of *VEGFA* (vascular endothelial growth factor-A) was downregulated (-1.17 FC; p=0.0102) and expression of its CDS was unchanged (1.01 FC; p=0.8334) (Fig. 66). The 3'UTR of *VEGFA*, a potent neuromodulator, undergoes a well-described binary switch to regulate its expression [113]. Our observations suggest the *VEGFA* 3' UTR undergoes an additional layer of regulation by shortening during collateral sprouting.

As our analysis with the GSE48611 and GSE72551 datasets show, reanalysis of publicly available datasets using updated annotations can yield additional information when compared to the use of the original CDFs. In our case, the region-based CDFs allow for a better understanding of 3' UTR dynamics through the reanalysis of publicly available data. While current high-throughput sequencing technologies may allow for a more complete picture, this custom CDF approach will allow for deeper insight with only minimal computational cost, taking advantage of the high volume of publicly available GeneChip® data.

Figure 66: Probe set expression levels within the gene, exon, and 3' UTR regions for 12 different genes

CHAPTER 9

CONCLUSIONS AND FUTURE DIRECTIONS

Microarrays were designed to examine the gene expression at the level of exons and genes but they do not examine the genes at the untranslated regions (UTRs). Additionally, they do not supply gene expression on transcript levels. Several researchers have called attention to the importance of getting expression levels for a specific transcript and created their own Chip Description Files (CDFs) to redesign probe sets. However, to this point in time analyses at the level of UTRs and individual exons has been minimal. It has been revealed that the untranslated regions of mRNA can dramatically change the regulation of protein and result in abnormal cell functions later on diseases. For that reason, achieving genome based screening of publicly available microarray data at the level of gene regions is important. Moreover, one of the limitations of microarray technologies is the design of probes based on available sequence and annotation data at the time of design. Based on our analysis, the percentage of uniquely mapping probes varies from 84% (rat) to 87% (human), indicating that changing knowledge about the genome itself plays a role in probe utilization. The significant differences between old and new genome assemblies and annotations make it necessary to update probe-gene targeting according to current knowledge to get more accurate interpretations from experimental results. In order to make it these possible, we created new CDFs by reassigning probes at the level of UTRS and individual exons. Moreover, we developed a tool for dynamically creating custom CDFs

based on the user's interest. We believe that our work will help researchers to reexamine already analyzed data in a different way

9.1 Summary of Contributions

- We developed an Affymetrix® GeneChip® probe remapping protocol at the level of genes, transcripts, untranslated regions (UTRs) and individual exons based on the latest genome (hg38, mm10, rn6) and Ensembl annotations for human, mouse, and rat. Our results illustrated how this framework affects the detection of differentially expressed genes, particularly when focusing on functional regions of interest. Removing probes that do not perfectly align to the genome or align to multiple locations can help to reduce false-positive differential expression, as can removal of probes in regions overlapping multiple genes.
- The main motivation of our work was profiling the contribution of UTR and exon regions to the gene expression levels globally. Our results indicate that features differentially expressed in either the gene-based or region-based CDF shows high concordance and separating out into functional regions allows for the detection of subtle changes in transcript formation.
- Our method can detect changes that would have been missed if the analysis was not separated into functional regions. For example, as shown in fig. 66 the 3' UTR of the *VEGFA* gene upregulates between day zero to day seven, however the exon regions of the *VEGFA* does not show a significant change.
- We developed an R package, *affyCustomCdf*, for allowing users to dynamically create custom CDFs. This is important since genomic knowledge grows and improves very fast. For example, when we performed analyses with our custom CDFs, the latest

Ensembl version was 85. Currently, Ensembl version 92 is available. Through the dynamic nature of our tool `affyCustomCdf`, user can easily benefit from the improvement over genomic knowledge and keep custom CDFs updated.

- Our custom CDF approach allows for deeper insight with only minimal computational cost, taking advantage of the high volume of publicly available GeneChip® data.

9.2 Post Dissertation Work

9.2.1 Dealing with Dynamic Biological Initializers

Some of the probe sequences may contain single nucleotide polymorphisms (SNPs) or may map to gene regions that produce variations of RNA products through a molecular process called RNA editing. It would be informative to query these probes and inform scientists about possible RNA editing effects within their analysis. These probes can be detected by mapping probes to the RNA editing site and SNP databases such as dbSNP. After identification of such probes, a special naming pattern could be used to mark probe sets containing these probes. Moreover, alternative probe sets can be created via modifying probes based on known RNA editing sites and SNPs and placed into custom CDFs.

9.2.2 Conservation of Gene Annotation

Some species or some genes of a species might be studied and annotated less than others. The number of annotated regions in a genome effect the number of probes used in a CDF because the probes that align to an unannotated region are going to be discarded in the proposed method. As seen in table 24 the number of probes annotated to a specific region is higher for the human genome which is well studied while there is likely to be differences in the number and construction of the gene between human and rat, the

transcriptomes are still highly similar. The most likely reason for the difference is incomplete annotation for rat.

TABLE 24: The number of probes per region and transcript after the annotation step

	UTR	CDS	Exon	Transcript
Human	268,863	108,960	440,252	2,129,643
Rat	100,774	56,425	163,498	275,588

In order to overcome the issue and gain discarded probes back, one can provide annotations to nonannotated regions of conserved genes across species by converting genome coordinates and genome annotations between two or more species.

For example, Fig. 67 shows the alignment of Affymetrix® Rat 230 2.0 probes that targets the *Camk4* gene. The probes (ending in _at) do not align to an annotated region of the rat and human genomes, but the mouse genome has a long UTR region annotation. If one can use annotations obtained from the mouse genome and apply them to the rat and human genomes, more probes can be added into probe sets.

Figure 67: Genome browser view of the gene *Camk4* with probes

Fig. 68 shows the alignment of Affymetrix® Rat 230 2.0 probes that target the *Otud7a* gene. In this case, both the human and mouse genomes have annotations for the probes. When the probes are aligned to the rat genome, they do not match to an annotation. In this example, mouse and human annotations could be used to annotate these probes.

Figure 68: Genome browser view of the gene *Otud7a* with probes

One way to accomplish this mapping is:

- I. Retrieve the annotation file for a species from the genome browser.
- II. Convert the annotation coordinates to another genome via the Lift Genome Annotations tool provided by the UCSC Genome browser.
- III. Use new annotations during the probe to annotation alignment.

9.2.3 Disease Drug Network Establishment Based on Genomic Expression Profiles

UTR region sequence and structure can dramatically change the regulation of a protein and result in abnormal cell functions and diseases. Our region based CDFs can be used to monitor the changes of UTR regions with different disease states and drug effects via analyses of whole-genome gene expression data available in the public repositories. The obtained expression profile similarities of diseases and drugs can be used to generate large scale disease-disease, drug-drug and disease-drug networks. A similar work has been done by Hu et al. They collected GEO datasets that is biologically and statistically comparable samples produced via same Affymetrix® GeneChip®. To avoid having multiple probe sets for a gene, they blasted probes of each probe set to NCBI RefSeq database. Based on the average distance of eleven probes to the 3' end, they classified probe sets and picked one probe set for each gene. Their first preference for the probe set selection is the one located on the 3' UTR, and they continued selection based on the shortest distance to the 3' UTR. Via analyses they generated disease and drug expression profiles from GEO

datasets and generated a large scale disease- drug network. They point out that the network can help finding new targets to existing drugs in a cost efficient way and drug target/pathway identification. Similar work can be done via our region based custom CDF for only untranslated regions of gene to reveal more information about UTRs, diseases and drugs relationship.

9.2.4 Untranslated Regions Gene Expression Profiling

Microarrays profile the relative differences in gene expression rather than absolute gene expression detection. Only genes that significantly vary between samples within an experiment are detected and the rest of the genes are ignored. Because of unknown and variable sensitivity of each probe set, microarrays do not guarantee that every probe set has the same dynamic range of gene expression values. As a result of it, comparison made between genes with different dynamic ranges does not supply accurate results. Creating a common gene expression profiling from publicly available microarray data via meta-analyses can provide an advantage for more accurate gene expression proofing. Such as a dynamic range interval can be calculated for each probe set of a specific microarray platform and the new experiment results can be judged based on the common reference. Creating common reference to normalize large amount of microarray data was already developed by many different groups. For example, Seita et al. [48] created an open platform for Affymetrix® mouse genome and made it available via a web site (<https://gexc.stanford.edu/>). But none of them analyzed the microarray data based on the regions of genes. One can use our region based custom CDFs file to achieve common gene expression profiling for untranslated regions of genes which can give a big advantage for future experiment design and analyses.

REFERENCES

1. Baldi, P. and G.W. Hatfield, *DNA microarrays and gene expression*. 2002, Cambridge, UK ; New York, NY: Cambridge University Press. xvi, 213 p.
2. Liu, G., et al., *NetAffx: Affymetrix probesets and annotations*. Nucleic Acids Research, 2003. 31.
3. Flight, R.M., A.M. Eteleeb, and E.C. Rouchka, *Affymetrix® Mismatch (MM) Probes: Useful after All*, in *Proceedings of the 2012 ASE/IEEE International Conference on BioMedical Computing*. 2012, IEEE Computer Society. p. 6-13.
4. Rouchka, E.C., A.W. Phatak, and A.V. Singh, *Effect of single nucleotide polymorphisms on Affymetrix match-mismatch probe pairs*. Bioinformation, 2008. 2(9): p. 405-11.
5. Rouchka, E.C., et al. *Computational Prediction of Genes Translationally Regulated by Cytoplasmic Polyadenylation Elements*. 2009. Berlin, Heidelberg: Springer Berlin Heidelberg.
6. Hilgers, V., et al., *Neural-specific elongation of 3' UTRs during Drosophila development*. Proceedings of the National Academy of Sciences, 2011. 108(38): p. 15864.
7. Ji, Z., et al., *Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development*. Proceedings of the National Academy of Sciences, 2009. 106(17): p. 7028.
8. Kuersten, S. and E.B. Goodwin, *The power of the 3' UTR: translational control and development*. Nature Reviews Genetics, 2003. 4: p. 626.
9. Revil, T., et al., *Alternative splicing is frequent during early embryonic development in mouse*. BMC Genomics, 2010. 11(1): p. 399.
10. Thomsen, S., et al., *Developmental RNA processing of 3'UTRs in Hox mRNAs as a context-dependent mechanism modulating visibility to microRNAs*. Development, 2010. 137(17): p. 2951.
11. Mercer, T.R., et al., *Targeted RNA sequencing reveals the deep complexity of the human transcriptome*. Nat Biotechnol, 2011. 30(1): p. 99-104.
12. Harrison, B.J., et al., *IB4-binding sensory neurons in the adult rat express a novel 3' UTR-extended isoform of CaMK4 that is associated with its localization to axons*. Journal of Comparative Neurology, 2014. 522(2): p. 308-336.
13. Harrison, B.J., et al., *Transcriptional changes in sensory ganglia associated with primary afferent axon collateral sprouting in spared dermatome model*. Genomics data, 2015. 6: p. 249-52.
14. Jansen, R.-P., *mRNA localization: message on the move*. Nature Reviews Molecular Cell Biology, 2001. 2: p. 247.
15. Prakash, N., et al., *Dendritic Localization of Rat Vasopressin mRNA: Ultrastructural Analysis and Mapping of Targeting Elements*. Vol. 9. 1997. 523-32.

16. Willis, D.E., et al., *Axonal localization of transgene mRNA in mature PNS and CNS neurons*. The Journal of neuroscience : the official journal of the Society for Neuroscience, 2011. 31(41): p. 14481-14487.
17. Akman, H.B., et al., *3'UTR shortening and EGF signaling: implications for breast cancer*. Human Molecular Genetics, 2015. 24(24): p. 6910-6920.
18. Fu, Y., et al., *Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing*. Genome Research, 2011. 21(5): p. 741-747.
19. Wang, L., et al., *The 3'UTR signature defines a highly metastatic subgroup of triple-negative breast cancer*. Oncotarget, 2016. 7(37): p. 59834-59844.
20. Polyadenylation. [cited 2018; Available from: <http://en.wikipedia.org/wiki/Polyadenylation>.
21. Derti, A., et al., *A quantitative atlas of polyadenylation in five mammals*. Vol. 22. 2012. 1173-83.
22. Curinha, A., et al., *Implications of polyadenylation in health and disease*. Nucleus (Austin, Tex.), 2014. 5(6): p. 508-519.
23. An, J.J., et al., *Distinct role of long 3' UTR BDNF mRNA in spine morphology and synaptic plasticity in hippocampal neurons*. Cell, 2008. 134(1): p. 175-87.
24. Lembo, A., F. Di Cunto, and P. Provero, *Shortening of 3'UTRs Correlates with Poor Prognosis in Breast and Lung Cancer*. PLOS ONE, 2012. 7(2): p. e31129.
25. Chatterjee, S. and J.K. Pal, *Role of 5'- and 3'-untranslated regions of mRNAs in human diseases*. Biology of the cell / under the auspices of the European Cell Biology Organization, 2009. 101(5): p. 251-62.
26. Barrett, L.W., S. Fletcher, and S.D. Wilton, *Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements*. Cellular and molecular life sciences : CMLS, 2012. 69(21): p. 3613-34.
27. Kratz, R.F., *Exploring the World of the Cell*, in *Molecular&Cell Biology for Dummies*. 2009, Wiley Publishing, Inc: Indianapolis. p. 9-14.
28. Splettstoesser, T. *File:Chromosome-DNA-gene.png*. 2015 [cited 2018; Available from: <https://commons.wikimedia.org/wiki/File:Chromosome-DNA-gene.png>.
29. Watson JD, C.F. *Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid*. 1953. Nature.
30. College, O. *File:229 Nucleotides-01.jpg*. 2013 [cited 2018; Available from: https://commons.wikimedia.org/wiki/File:229_Nucleotides-01.jpg.
31. CRICK, F.H.C. *ON PROTEIN SYNTHESIS*. 1958.
32. Yikrazuul. *File:The tRNA cloverleaf general.svg*. 2010; Available from: https://commons.wikimedia.org/wiki/File:The_tRNA_cloverleaf_general.svg.
33. *Chapter 1: How Genes Work*. The New Genetics [cited 2015 11/5]; Available from: <http://publications.nigms.nih.gov/thenewgenetics/chapter1.html>.
34. Layer, R.M., et al., *Binary Interval Search: a scalable algorithm for counting interval intersections*. Bioinformatics, 2013. 29(1): p. 1-7.
35. Bowtie. *Bowtie:Manual*. [cited 2015 December]; Available from: <http://bowtie-bio.sourceforge.net/manual.shtml>.
36. hannonlab. *FASTX-Toolkit*. 2015; Available from: http://hannonlab.cshl.edu/fastx_toolkit/index.html.

37. *Noncoding DNA*. [cited 2018; Available from: http://en.wikipedia.org/wiki/Noncoding_DNA.
38. *Intron*. [cited 2018; Available from: <https://en.wikipedia.org/wiki/Intron>.
39. Wikipedia. *Central dogma of molecular biology*. Wikipedia 2018 [cited 2018; Available from: c.
40. Crick, F. *Ideas on Protein Synthesis*. 1956.
41. Ruiz, L.M. *File:DNA replication en.svg*. 2007 [cited 2018; Available from: https://commons.wikimedia.org/wiki/File:DNA_replication_en.svg.
42. Programme, G.E. *File:Process of transcription (13080846733).jpg*. 2014 [cited 2018; Available from: [https://commons.wikimedia.org/wiki/File:Process_of_transcription_\(13080846733\).jpg](https://commons.wikimedia.org/wiki/File:Process_of_transcription_(13080846733).jpg).
43. Calibuon. *File:Transcription.jpg*. 2008 [cited 2018; Available from: <https://commons.wikimedia.org/wiki/File:Transcription.jpg>.
44. Kratz, R.F., *Dead or Alive: Viruses*, in *Molecular & Cell Biology*. 2009, Wiley Publishing.
45. McCarthy, B.J. and J.J. Holland, *Denatured DNA as a direct template for in vitro protein synthesis*. Proceedings of the National Academy of Sciences of the United States of America, 1965. 54(3): p. 880-6.
46. Nirenberg, M.W., *The genetic code, Nobel Lecture*. 1968.
47. Gentleman, R.C., et al., *Bioconductor: open software development for computational biology and bioinformatics*. Genome biology, 2004. 5(10): p. R80.
48. Seita, J., et al., *Gene Expression Commons: an open platform for absolute gene expression profiling*. PloS one, 2012. 7(7): p. e40321.
49. Guttman, A., *R-trees: A Dynamic Index Structure for Spatial Searching*. SIGMOD Rec, 1984. 14: p. 47--57.
50. Southern, E.M., *Detection of specific sequences among DNA fragments separated by gel electrophoresis*. Journal of molecular biology, 1975. 98(3): p. 503-17.
51. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors*. Proceedings of the National Academy of Sciences of the United States of America, 1977. 74(12): p. 5463-7.
52. Alwine, J.C., D.J. Kemp, and G.R. Stark, *Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes*. Proceedings of the National Academy of Sciences of the United States of America, 1977. 74(12): p. 5350-4.
53. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome*. Genome biology, 2009. 10(3): p. R25.
54. Schena, M., et al., *Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray*. Science, 1995. 270(5235): p. 467-470.
55. Helen Causton, J.Q., *Alvis Brazma Microarray Gene Expression Data Analysis: A Beginner's Guide*. Wile.
56. Knudsen, S., *Guide to analysis of DNA microarray data*. 2nd ed. 2004, Hoboken, N.J.: Wiley-Liss. xv, 168 p.
57. David B. Allison, G.P.P., T. Mark Beasley, Jode W. Edwards, *The Problem of Errors and Multiplicity Control*, in *DNA Microarrays and Related Genomics*

- Techniques: Design, Analysis, and Interpretation of Experiments*. 2005, CRC Press.
58. Scherer, A., *Microarray Platforms and Aspects of Experimental Variation*, in *Batch Effects and Noise in Microarray Experiments : Sources and Solutions*. 2009, Wiley.
  59. Miller, M.B. and Y.-W. Tang, *Basic Concepts of Microarrays and Potential Applications in Clinical Microbiology*. *Clinical Microbiology Reviews*, 2009. 22(4): p. 611-633.
  60. Affymetrix®, *GeneChip® Human Genome U133 Arrays Technical Note*. 2003-2007.
  61. Affymetrix®, *Design and Performance of the GeneChip® Human Genome U133 Plus 2.0 and Human Genome U133A 2.0 Arrays*. 2003.
  62. Su, A.I., et al., *A gene atlas of the mouse and human protein-encoding transcriptomes*. *Proceedings of the National Academy of Sciences of the United States of America*, 2004. 101(16): p. 6062-6067.
  63. Weick, J.P., et al., *Deficits in human trisomy 21 iPSCs and neurons*. *Proceedings of the National Academy of Sciences of the United States of America*, 2013. 110(24): p. 9962-7.
  64. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets—update*. *Nucleic Acids Research*, 2013. 41(Database issue): p. D991-D995.
  65. Liu, G., et al., *NetAffx: Affymetrix probesets and annotations*. *Nucleic acids research*, 2003. 31(1): p. 82-86.
  66. Wagner L, A.R., *UniGene*, in *The NCBI Handbook [Internet]. 2nd edition*. 2013 Nov 14, National Center for Biotechnology Information (US): Bethesda (MD).
  67. Maglott, D.R., et al., *NCBI's LocusLink and RefSeq*. *Nucleic Acids Research*, 2000. 28(1): p. 126-128.
  68. Ncbi Resource Coordinators, *Database resources of the National Center for Biotechnology Information*. *Nucleic Acids Research*, 2015. 43(Database issue): p. D6-D17.
  69. Chalifa-Caspi, V., et al., *GeneAnnot: comprehensive two-way linking between oligonucleotide array probesets and GeneCards genes*. *Bioinformatics*, 2004. 20(9): p. 1457-8.
  70. Tatusova, T., et al., *RefSeq microbial genomes database: new representation and annotation strategy*. *Nucleic Acids Research*, 2014. 42(Database issue): p. D553-D559.
  71. Rouchka, E.C., A.W. Phatak, and A.V. Singh, *Effect of single nucleotide polymorphisms on Affymetrix® match-mismatch probe pairs*. *Bioinformation*, 2008. 2(9): p. 405-411.
  72. Benovoy, D., T. Kwan, and J. Majewski, *Effect of polymorphisms within probe-target sequences on oligonucleotide microarray experiments*. *Nucleic acids research*, 2008. 36(13): p. 4417-4423.
  73. Mohammad, F., et al., *AbsIDconvert: An absolute approach for converting genetic identifiers at different granularities*. *BMC bioinformatics*, 2012. 13: p. 229-229.
  74. Gautier, L., et al., *Alternative mapping of probes to genes for Affymetrix chips*. *BMC bioinformatics*, 2004. 5: p. 111.

75. Dai, M., et al., *Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data*. Nucleic Acids Research, 2005. 33(20): p. e175-e175.
76. Harbig, J., R. Sprinkle, and S.A. Enkemann, *A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array*. Nucleic Acids Research, 2005. 33(3): p. e31.
77. Liu, H., et al., *AffyProbeMiner: a web resource for computing or retrieving accurately redefined Affymetrix probe sets*. Bioinformatics, 2007. 23(18): p. 2385-90.
78. Lu, J., et al., *Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: high-resolution annotation for microarrays*. BMC bioinformatics, 2007. 8: p. 108.
79. Ferrari, F., et al., *Novel definition files for human GeneChips based on GeneAnnot*. BMC bioinformatics, 2007. 8: p. 446.
80. Yu, H., et al., *Transcript-level annotation of Affymetrix probesets improves the interpretation of gene expression data*. BMC bioinformatics, 2007. 8: p. 194.
81. de Leeuw, W.C., et al., *Salvaging Affymetrix probes after probe-level re-annotation*. BMC research notes, 2008. 1: p. 66.
82. Ballester, B., et al., *Consistent annotation of gene expression arrays*. BMC genomics, 2010. 11: p. 294.
83. Risueno, A., et al., *GATEXplorer: genomic and transcriptomic explorer; mapping expression probes to gene loci, transcripts, exons and ncRNAs*. BMC bioinformatics, 2010. 11: p. 221.
84. Yin, J., et al., *Integrating multiple genome annotation databases improves the interpretation of microarray gene expression data*. BMC genomics, 2010. 11: p. 50.
85. Benson, D.A., et al., *GenBank*. Nucleic Acids Research, 2012. 40(Database issue): p. D48-D53.
86. Cunningham, F., et al., *Ensembl 2015*. Nucleic Acids Research, 2015. 43(D1): p. D662-D669.
87. Kent, W.J., *BLAT—The BLAST-Like Alignment Tool*. Genome research, 2002. 12(4): p. 656-664.
88. Safran, M., et al., *GeneCards Version 3: the human gene integrator*. Database, 2010. 2010.
89. Kent, W.J., et al., *The Human Genome Browser at UCSC*. Genome Research, 2002. 12(6): p. 996-1006.
90. Huber, W. and R. Gentleman, *matchprobes: a Bioconductor package for the sequence-matching of microarray probe elements*. Bioinformatics, 2004. 20(10): p. 1651-1652.
91. Mighell, A.J., A.F. Markham, and P.A. Robinson, *Alu sequences*. FEBS Letters, 1997. 417(1): p. 1-5.
92. Maglott, D., et al., *Entrez Gene: gene-centered information at NCBI*. Nucleic Acids Research, 2011. 39(Database issue): p. D52-7.
93. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data*. Biostatistics, 2003. 4(2): p. 249-64.

94. Altschul, S.F., et al., *Basic local alignment search tool*. Journal of molecular biology, 1990. 215(3): p. 403-10.
95. Thierry-Mieg, D. and J. Thierry-Mieg, *AceView: a comprehensive cDNA-supported gene and transcripts annotation*. Genome biology, 2006. 7(Suppl 1): p. S12-S12.
96. *Latin Square Data for Expression Algorithm Assessment*. Available from: http://www.affymetrix.com/support/technical/sample_data/datasets.affx.
97. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data*. Biostatistics, 2003. 4(2): p. 249-264.
98. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response*. Proceedings of the National Academy of Sciences of the United States of America, 2001. 98(9): p. 5116-5121.
99. Kersey, P.J., et al., *The International Protein Index: an integrated database for proteomics experiments*. Proteomics, 2004. 4(7): p. 1985-8.
100. Slater, G.S.C. and E. Birney, *Automated generation of heuristics for biological sequence comparison*. BMC bioinformatics, 2005. 6: p. 31-31.
101. Kasprzyk, A., et al., *Ensembl: A Generic System for Fast and Flexible Access to Biological Data*. Genome Research, 2004. 14(1): p. 160-169.
102. Pang, K.C., et al., *RNAdb 2.0--an expanded database of mammalian non-coding RNAs*. Nucleic acids research, 2007. 35(Database issue): p. D178-82.
103. Sprague, J., et al., *The Zebrafish Information Network (ZFIN): the zebrafish model organism database*. Nucleic acids research, 2003. 31(1): p. 241-243.
104. Flicek, P., et al., *Ensembl 2008*. Nucleic Acids Research, 2008. 36(Database issue): p. D707-14.
105. Michael Love , S.A., Wolfgang Huber *RNA-seq workflow at the gene level*.
106. Ferragina, P. and G. Manzini, *Opportunistic data structures with applications*, in *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*. 2000, IEEE Computer Society. p. 390.
107. Thomas H. Cormen; Charles E Leiserson.; Ronald L. Rivest, C.S., *Introduction to Algorithms*. Second ed. 2002.
108. Alekseyenko, A.V. and C.J. Lee, *Nested Containment List (NCList): a new algorithm for accelerating interval query of genome alignment and interval databases*. Bioinformatics, 2007. 23(11): p. 1386-93.
109. Li, J., et al., *Normalization, testing, and false discovery rate estimation for RNA-sequencing data*. Biostatistics, 2012. 13(3): p. 523-38.
110. Harrison, B.J., et al., *IB4-binding sensory neurons in the adult rat express a novel 3' UTR-extended isoform of CaMK4 that is associated with its localization to axons*. J Comp Neurol, 2014. 522(2): p. 308-36.
111. Fernandes, H.B., et al., *High-Affinity Kainate Receptor Subunits Are Necessary for Ionotropic but Not Metabotropic Signaling*. Neuron. 63(6): p. 818-829.
112. Pickard, B.S., et al., *A common variant in the 3'UTR of the GRIK4 glutamate receptor gene affects transcript abundance and protects against bipolar disorder*. Proc Natl Acad Sci U S A, 2008. 105(39): p. 14940-5.
113. Ray, P.S., et al., *A stress-responsive RNA switch regulates VEGFA expression*. Nature, 2009. 457(7231): p. 915-9.

APPENDIX A

AffyCustomCdf PARAMETERS

A.1 Original CDF

The original CDF of the selected Affymetrix® GeneChip® microarray must be supplied to the tool by the user. It is used as a source to obtain control probes and PM and MM tbase and pbase probe information. It is taken via `originalCdfName` parameter. It can be obtained from Affymetrix® NetAffx™ or GEO. As an example, the original CDF file for Rat 230 2.0 is included inside the scripts directory of the github repository branch `affyCustomCdfFull`.

(<https://github.com/UofLBioinformatics/affyCustomCDF/tree/affyCustomCdfFull>)

A.2 Probe Alignment File

The `affyCustomCdf` expects probe mapping as a text file. The probe mapping file consists of one line per probe, each containing five columns of data. Columns must be tab delimited without a header line. Columns are in the following order: x location of a probe, y location of a probe, sense/antisense of a probe (-/+), chromosome name and chromosomal starting location. Fig. 69 demonstrates first ten lines of the Rat 230 2.0 microarray probe mapping file. The file is included inside the scripts directory of the github repository branch `affyCustomCdfFull`.

335	337	-	8	22161426
99	499	-	8	22161407
726	583	-	8	22161386
386	415	-	8	22161292
101	649	-	8	22161280
180	413	-	8	22161263
13	31	-	8	22161145
258	663	-	8	22160989
392	57	-	8	22160974

Figure 69: First ten lines of the Rat 230 2.0 microarray probe mapping file

A.2.1 Recommended probe mapping text file creation

The DNA sequences for perfect match (PM) probes is obtained from the Affymetrix® Netaffx™ web site in a FASTA file format. PM probes are aligned to the genome using Bowtie version 1.0.1 with the parameters -v 0 and -m 1 which returns the alignment results for the probes that align to one genomic location with 100% identity. Also, the suppress parameter can be used to eliminate some output for clarity. The following command line is an example for aligning Rat 230 2.0 PM to Rat assembly rn6. It suppresses the 5th, 6th and 7th default Bowtie outputs.

```
$ ./bowtie -v 0 -best -m 1 rnRnor6Chromosome -f Rat230_2.fa rnRnorchromosome6.txt
--suppress 5,6,7 --quiet --max maxrnRnorChromosome6.txt
--un unalignedrnRnorChromosome6.txt
```

Figure 70: Bowtie command line for probe mapping

If the index of the intended genome of an organism is not supplied by Bowtie, the assembled unmasked genomic DNA sequences of every chromosome can be downloaded from one of the online repositories such as Ensembl and NCBI and indexes can be created from DNA sequences via Bowtie-Build version 1.0.1 with default parameters. Fig. 71 show the one line of mapping result file.

```
Rat230_2:1367453_at; 335; 337; 1100; Antisense; - 8 22161426
```

Figure 71: Mapping result of Bowtie for Rat 230 2. microarray

The file can be formatted via an R script to obtain the proper column arrangement (Fig. 72). More script examples and data files can be found under the scripts folder of the github repository branch `affyCustomCdfFull` with the file name `ProbeFileClear.R`.

```
fileName = "rat.txt"
probes = read.csv2(fileName, header = FALSE, sep = "\t", quote = "\"",
dec = ",", fill = TRUE, col.names = c("names","sense","chromosome", "start",
"empty") ,comment.char = "",stringsAsFactors = FALSE)

X = unlist(lapply(strsplit(probes[,1],";"), "[[" , 2))
Y = unlist(lapply(strsplit(probes[,1],";"), "[[" , 3))

probeTable = data.table(X = as.numeric(X), Y = as.numeric(Y),
direction=unlist(probes$sense), chromosome=unlist(probes$chromosome),
start=unlist(probes$start))

write.table(probeTable, "Rat230-2Rn6ChrAligned.txt", quote = FALSE,
row.names = FALSE, col.names = FALSE, sep="\t")
remove(probeTable,X,Y,probes)
```

Figure 72: Example R code for clearing Bowtie mapping results

A.3 Obtaining annotation file

The `affyCustomCdf` tool accepts annotations as a General/Gene Transfer Format (GTF). In the GTF file, fields must be tab delimited. The GTF format consist of nine columns of data per feature. The columns are seqname, source, feature, start, end, score, strand, frame and attribute. Comment lines in the file must start with `#`. The feature column is being used to classify probes for gene features. Therefore, region features must contain transcript, EXON, CDS and UTR key words (not case sensitive). If separation between 3' UTR and 5' UTR is desired, features must include the direction information such as `three_prime utr` or `five_prime utr`. Otherwise annotation will be performed without

classification of 3' and 5'. For more information please check the Ensembl GFF/GTF file format definition. <http://www.ensembl.org/info/website/upload/gff.html>. A sample GTF file can be found in the scripts directory of the github repository branch `affyCustomCdfFull`.

A.4 Selecting Parameters

The `affyCustomCdf` tool takes thirteen parameters. This includes the original CDF (Section A.1), probe alignment file (Section A.2) and annotation (GTF) file (Section A.3) which must be supplied by the user. The remaining parameters are optional, with default value set. The parameters are:

- **originalCdfName:** String type. The original CDF file of the selected Affymetrix® GeneChip® technology. It can be obtained from Affymetrix® NetAffx™ web site. For example, `Rat230_2.cdf` file for the Rat 230 2.0 microarray.
- **probeAlignmentFile:** String type. The tab separated probe alignment file. Please see section A.1 for more details.
- **gtfFileName:** String type. The General/Gene Transfer Format (GTF) file name. Please see section A.3 for more details.
- **newCDFName:** String type. Name of the created custom CDF. If the user does not provide a name, a name will be created based on the template provided in Fig. 73.
- **reportFile:** String type. Name of the report file name. If the user does not provide a name, a name will be created based on the template `report_originalCdfName.txt`.


```

originalCdfName_type_min_minProbeSetNumber_D_SD.cdf
* type the value of the type parameter
* min stands for minimum
* minProbeSetNumber indicates the minimum number of probe per probe set
* D stands for direction
* SD indicates the value of direction parameter

```

Figure 73: File naming for custom CDFs

- **controlProbeSetNumber:** Number type. The number of control probe sets in the original CDF. Control probe sets were being designed to check the quality of the experiment. They can be detected in the CDF based on the probe set names which usually starts with AFFX prefix such as AFFX_ratb2/X14115_at. If the number of control probe set is given, they will be included in the custom CDF without changes otherwise they will not be included. The default value is 0 (not included).
- **minProbeSetNumber:** Number type. The minimum number of probes in a probe set. Probe sets with less than the minimum number will not be included in the custom CDF. The default value is three.
- **probeLength:** Number type. Number of bases in a probe. In most cases length of a probe is twenty-five. For different sizes, it can be obtained via checking a sequence of a probe belongs to an Affymetrix® GeneChip®. The default value is 25.
- **SD:** Number type. Sense/antisense relationship between probes and annotations. Possible values are 0 (no direction), 1 (same direction) and any value except 0 and 1 (opposite direction). The default value is 1 (Same direction).
  - When SD is 0, direction will not be considered during probes to annotations mapping.

- When SD is 1, mapping is performed between same directions such as the sense probes are being mapped to the sense annotations.
- When SD is any value except 0 and 1, mapping is performed between opposite directions such as the sense probes are being mapped to the anti-sense annotations.
- **type = “regionG”:** String type. The type of the created custom CDF. Options are regionG, gene, transcript. Default value is regionG. G
  - regionG: When regionG option is selected, probe sets are designed to target a specific region (exon, UTR) of a gene and consist of probes which map to the same region of a gene.
  - gene: When regionG option is selected, probe sets are designed to target genes and consist of probes which map to the same gene.
  - transcript: When transcript option is selected, probe sets are designed to target transcripts and consist of probes which map to the same transcript.
- **transcriptShare:** Bool type. It defines whether to allow probe sharing between transcripts of a gene. It can only be used when transcript type is selected. Possible values are FALSE for not allowing probe share and TRUE for allowing probe share. The default value is FALSE.
- **junction:** Bool type. It defines whether we are adding probes to probe sets that map onto a junction of a specific region of a gene. It can be used when regionG or gene option is selected. Options are FALSE for no junction and TRUE to allow junctions. The default value is FALSE.

- **uniqueProbe:** Bool type. It defines whether to use probes that map onto more than one annotations. It is an option for the user to examine unique probes to one specific annotation. It can only be used with the regionG type. Possible values are FALSE for not unique, TRUE for unique. The default value is FALSE.

A.5 Running The affyCustomCdf Tool

To create a custom CDF via affyCustomCdf tool, one must call the createAffyCustomCdf function after installation of the tool. Fig. 74 shows example R scripts.

```
#Creates region based CDF for the Rat 230 2.0 with default values
createAffyCustomCdf("Rat230_2.cdf","rat230Probes.txt",
"Rattus_norvegicus.Rnor_6.0.85.gtf")

#Creates gene based CDF for the Rat 230 2.0
createAffyCustomCdf("Rat230_2.cdf","rat230Probes.txt",
"Rattus_norvegicus.Rnor_6.0.85.gtf",
controlProbeSetNumber = 57, minProbeSetNumber = 2, type = "gene")

#Creates transcript based CDF for the HG-133 Plus 2
createAffyCustomCdf("HG-U133_Plus_2.cdf","human133Probes.txt",
"Homo_sapiens.GRCh38.77.gtf", controlProbeSetNumber = 62, minProbeSetNumber = 2,
type = "transcript")
```

Figure 74: Calling createAffyCustomCdf functio

CURRICULUM VITAE

NAME Ernur Saka

ADDRESS Department of Computer Science and Computer Engineering
University of Louisville
Louisville, KY 40292

EDUCATION Master of Science in Computer Science (MSCS)
University of Louisville
B.S. in Computer Engineering
University of Dokuz Eylül

EXPERIENCE University of Louisville Teaching Assistant
Intel Graphics Software Engineer Intern
University of Louisville Research Assistant
Birim Information Operation Consultancy and Commerce Limited
Company Turkey Engineering Intern
Turkey Telecommunications Authority Telecommunication Board
Engineering Intern
University Of Dokuz Eylül Hardware Intern

AWARDS

CECS Arthur M. Riehl Award for excellent academic performance and contributions to departmental activities 2018

Scholarship sponsored by the National Science Foundation and Adobe Systems Inc. to attend the 2012 and the 2011 Grace Hopper Celebration of Women in Computing 2012,2011

IEEE International Conference on Computational Advances in Bio and Medical Sciences ICCABS Travel Award 2016, 2012

University of Louisville Sponsored Research Tuition Award for PhD 2011

Acceptance to the SIGGRAPH as a Student Volunteer from a worldwide application pool and representing UofL 2010

PUBLICATIONS

Ernur Saka, Benjamin J. Harrison, Kirk West, Jeffrey C. Petruska, Eric C. Rouchka Journal Paper BMC Genomics 18(Suppl 10):875: Framework for reanalysis of publicly available affymetrix® geneChip® data sets based on functional regions of interest

Ernur Saka, Benjamin J. Harrison, Kirk West, Jeffrey C. Petruska, Eric C. Rouchka Region-based custom chip description formats for reanalysis of publicly available Affymetrix® genechip® data sets Computational Advances in Bio and Medical Sciences (ICCABS), 2016 IEEE 6th International Conference on 1-Jan 2016 IEEE

Ernur Saka, Eric Rouchka Poster Paper: Image Registration and Visualization Tool For In-Situ Gene Expression Images, IEEE ICCABS