Estimating Cell-Type Profiles and Cell-Type Proportions in Heterogeneous Gene Expression Data

by

David A. Pritchard

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
August 4, 2012

Keywords: eQTL, heterogeneous, mouse brain

Approved by

Peng Zeng, Chair, Associate Professor of Mathematics and Statistics
Mark Carpenter, Professor of Mathematics and Statistics
Nedret Billor, Associate Professor of Mathematics and Statistics
George Flowers, Dean of the Graduate School and Professor of Mechanical Engineering

Abstract

Understanding the mechanisms underlying natural variation in gene expression is an important question in medical and evolutionary genetics. Many studies intend to compare either *(i)* cell-type expression profiles across individuals for the same cell-type, or *(ii)* cell-type expression profiles within an individual for different cell-types (NIH 2012). Naturally, accurate estimates of these expression profiles is of great importance. However, the presence of heterogeneity of cell-types in gene expression data can result in inaccurate estimates of such cell-type expression profiles (Leek and Storey 2007).

The standard statistical method for assaying gene expression data is to use a simple linear regression model, with the assumption that the presence of minor alleles in the genotype has an additive effect on gene expression levels (Veyrieras 2008). This method assumes that the observed gene expression data has a homogeneous composition of a single cell-type. However there are many scenarios where it may be more appropriate to assume that observed gene expression data is composed of two cell-types; for example a brain tissue sample would presumably have a heterogeneous mixture of neuron and glial cell-types (GeneNetwork 2012).

Previous studies have developed methodologies for estimating cell-type expression profiles given prior information regarding individual cell-type proportions; or conversely for estimating cell-type proportions with prior knowledge of cell-type expression profiles. This thesis derives a computational method for estimation of both the cell-type expression profiles and individual cell-type proportions for a two cell-type model, without any prior information. The parameter estimation techniques are based on an alternating-regression least-squares process. This methodology is applied to both simulated data and a real dataset, and the results are examined.

Acknowledgments

I would like to thank the members of my committee, Dr. Peng Zeng, Dr. Mark Carpenter, and Dr. Nedret Billor for their support of this work. I would also like to thank the other faculty and staff of the Auburn University Mathematics and Statistics department, who's hard work and dedication allow all of the students here the opportunity to grow and thrive.

This work would not have been possible without the support of my brother, Dr. Jonathon Pritchard, who provided the motivation for the project and suggested the problem. I would like to thank John Blischak for helping me get this project off the ground, and for tirelessly answering my biology questions. I would like to give a special thanks to Dr. Abe Palmer, who's understanding of and contributions to the biological aspects of this project have been invaluable.

Most importantly, I would like to again thank my advisor, Dr. Peng Zeng. Over the past two years Dr. Zeng has been both my professor in the classroom and my advisor on this project. He has easily been the single biggest influence on me as a scientist. I cannot say enough about how much I have grown under his tutelage. Thank you so much for everything.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Biotechnology plays an increasingly important role in the modern human experience. Biological applications have often been the catalyst for the development of statistics as a science in the 20th century. Whether it was Pearson and study of Darwin's hypothesis, Gosset's search to brew the perfect batch of Guiness, or Fisher's studies in crop variation, statistics have often been motivated by biological processes. But Fisher and company could not have imagined the wealth of biotechnological applications that pervade today's world.

In the 21st century, foods are grown and processed using cultivated microorganisms. Yeasts and mold are being studied as means to create proteins to supplement traditional foods as the planet's growing population strains natural resources. Genetic engineering has helped us create genetically modified plants that grow better and have more bountiful yields, and are more resistant to diseases and environmental hazards.

Biotechnology is increasingly used in the fields of medicine and pharmocology. Recombinant DNA is used to produce specific enzymes in the body. Hormones can be synthesized through genetic engineering. Antibiotics and vaccines have been our first line of defense against disease for close to 100 years; now scientists race to develop new treatments for increasingly resistant strains of disease. New vaccines are being studied to prevent illnesses such as hepatitis, malaria, and AIDS.

Some of the newest applications of biotechnology are in the field of bioinformatics. The abilities to sequence genetic data, RNA, amino acids and protein sequences have provided scientists a wealth of information from which to study the very building blocks of biological organisms. Studies in genetics could reveal the causal links behind diseases like cancer or diabetes, and help create treatments and cures for diseases (Dubey 2006).

The studies of bioinformatics, and in particular statistical genetics, are still in their relative infancies. New technologies allow us to study the underlying mechanisms through which biological organisms function. One of the predominant questions in genetics today is, what are the mechanisms through which genetic variation occurs? The research in this thesis provides a tool to gain insight into this incredibly important question.

1.1 Background

Deoxyribonucleic acid, commonly known as DNA, is the hereditary material of an organism. The information contained in DNA gives the organism the instructions which controls its functions. This genetic information is stored in the double-helix structure first proposed by Watson and Crick (1963). The double-helix is the composition of two strands of complementary chemical bases wound together to form a spiral helix shape. Each strand contains a string of chemical bases. There are a total of four different chemical bases: adenine, guanine, cytosine, and thymine. These bases are typically referred to using the abbreviations A, C, G, and T respectively. Each base is paired with its complement: A pairs with T, and C pairs with G. Thus, the information contained in DNA can be represented by just one of the two strands, since one strand is completely determined by the other. The chemical bases of the strands, together with binding sugar and phosphate molecules are altogether called nucleotides.

The DNA molecule is subdivided into tightly coiled structures called chromosomes. Contained within each of these chromosomes are functional units of hereditary information called genes. One of the primary mechanisms through which DNA regulates bodily activity is through the production of proteins. Proteins are molecules required for the structure, function, and regulation of the body's tissues and organs, and the amount and timing of production of proteins are what determines an individual's phenotype. An individual's phenotype is the composition of the physical embodiments or traits of the individual's genetic makeup. Information in each gene regulates the production of proteins in the body. Thus,

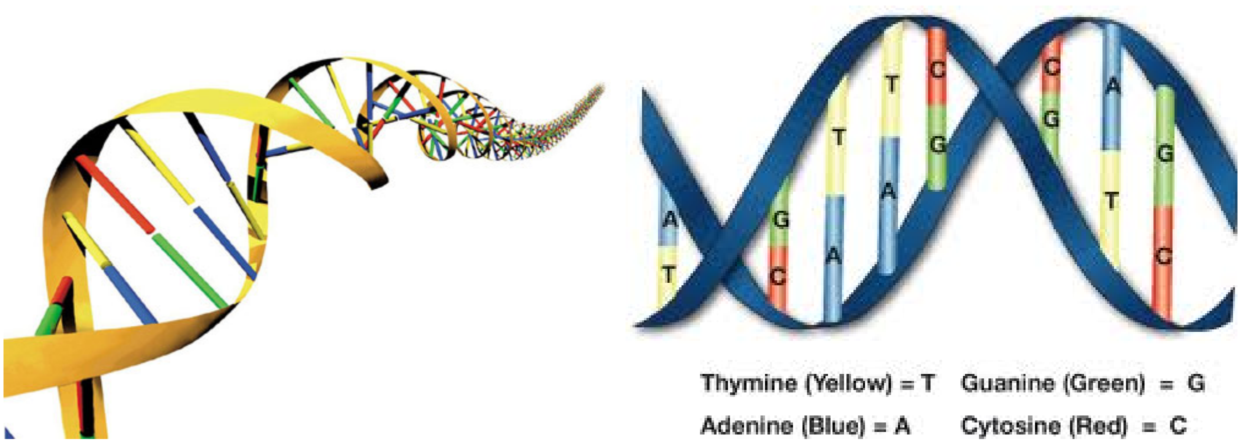


Figure 1.1: *DNA double helix*

this regulation of proteins is a key component of the study of genetics. The quantity of protein produced by a gene is referred to as gene expression.

Genetic variation refers to differences in the genetic code between individuals of the same species. The DNA of any two members of a species is exceedingly similar. In humans, for example, it is typical for over 99% of the genes of any two individuals to be identical. It is the differences in two individual's genetic code which accounts for different physical traits between the individuals. Since physical differences are largely the manifestation of differences in protein production, scientists wish to understand the relationship between genetic variation and protein production (Avila 1995).

Almost all locations in the genetic code are exceedingly stable across members of a population. However, geneticists have observed that a tiny proportion of locations in the genome have a preponderance of mutations across individuals. These locations are called single nucleotide polymorphisms (SNPs). If these locations are known, then a study of genetic variation between individuals can be reduced to studying these SNPs.

Consider the following illustration of a SNP. Suppose that a DNA segment is taken at the same location from two individuals contain the nucleotyde strings CAG$\underline{\text{T}}$AG and CAG$\underline{\text{A}}$AG. Then the two strings have a difference of a single nucleotide. We say that there are two alleles (forms of the nucleotyde): T, and A. Each individual has two sources of genetic

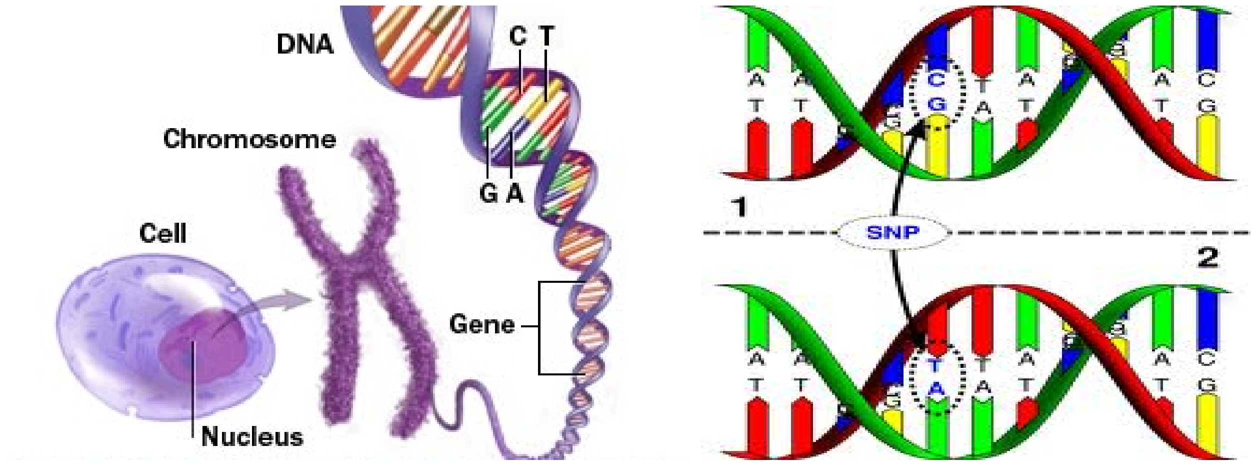
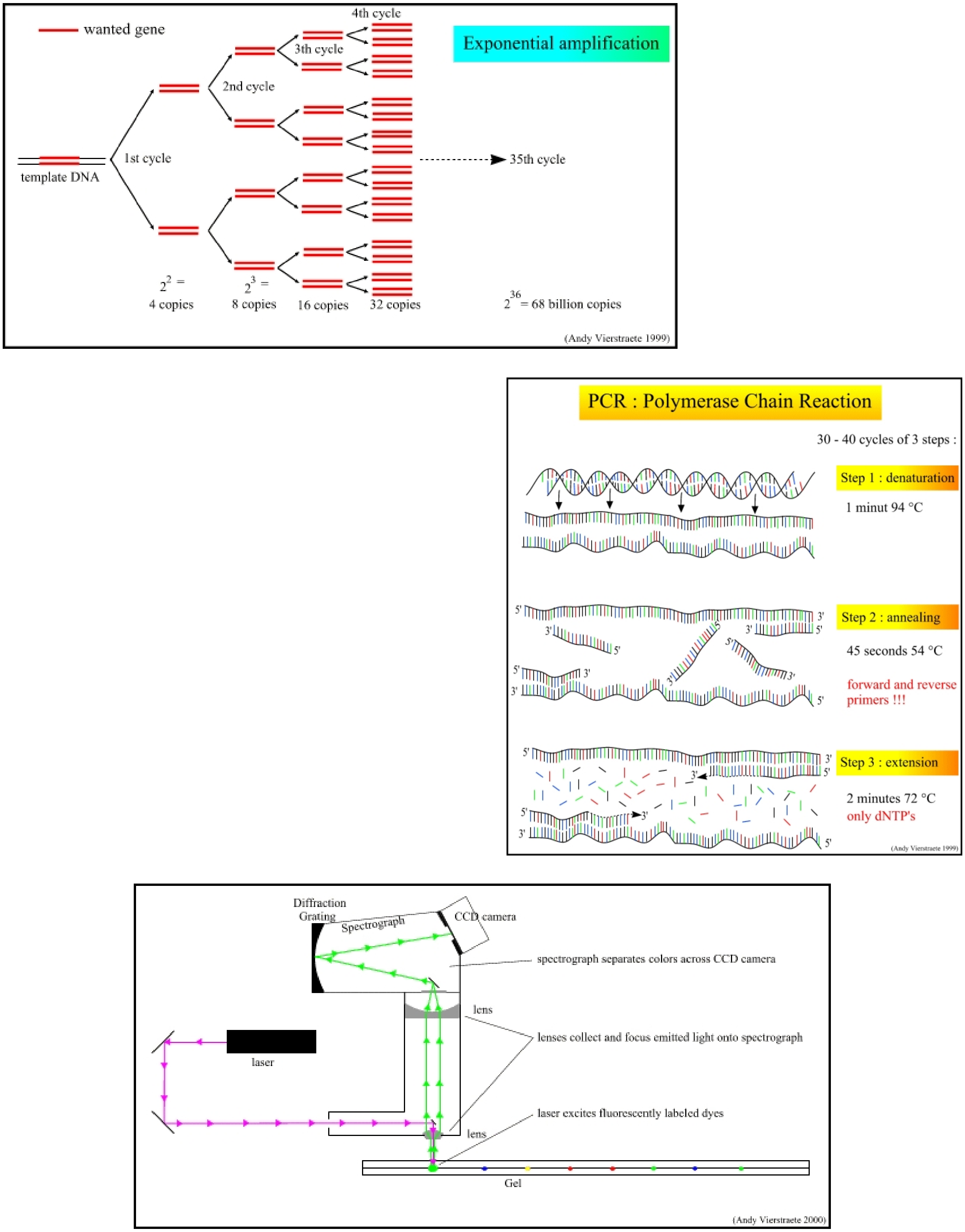


Figure 1.2: *Gene within a cell chromosome / SNP illustration*

code: one from the mother and one from the father. So going back to our example, assuming that there are only two alleles at the SNP, then an individual may have the following allele combinations: AA, AT, TA, or TT. Let g be the number of times T occurred at the SNP, then $g \in \{0, 1, 2\}$.

In order to assay the relationship between SNPs and gene expression, one must first begin with knowledge of the genotypes and the gene expression levels of a cell-type for a group of individuals. A genotype is a measure of how that individual's genomic sequence differs from a representative genomic sequence from the species (NIH 2012).

There are various methods of genotyping. We describe the method used to procure the data used by our study as an illustration (actual dataset discussed in detail later). This method began with the polymerase chain reaction (PCR) method of amplification described by Love *et al.* (1990) and Dietrich *et al.* (1992). The process involves taking a DNA segment and denaturing the strands, then synthetic DNA blocks called primers are exposed to the strands and bind to them wherever they can find a complementary match. The matched strands are then removed from the remaining primers. The process is repeated k times, so the DNA is amplified to a resulting 2^k DNA segments. Figure 1.3 provides a visual illustration of the process of genotyping.



(Andy Vierstraete 1999)

(Andy Vierstraete 1999)

(Andy Vierstraete 2000)

Figure 1.3: *Genotyping process*

The sequencing process begins by separating the DNA sequences into individual strands of DNA. The molecules used to separate the chains are labeled using dyes which emit light at different wavelengths. A laser sequencer then sends light through the ordered fragments to create a sequence of different colors which can be then be expressed as the nucleic acid bases. The strands are then pieced together to create the genomic sequence (Peirce 2004).

Scientists are often interested in measuring the quantity of proteins produced by a gene. However it is difficult to directly measure protein levels, so proxy measures of protein levels have been developed instead. These measures are chosen as a consequence of the biological processes from which proteins are produced.

The production of proteins is controlled by the genes through the processes of transcription and translation. The first step in protein synthesis, transcription, involves the production of messenger ribonucleic acid (mRNA). mRNA is a sequence of nucleotides created by unzipping a stretch of DNA, and creating a strand of mRNA complementary to the one of the DNA pairs.

Translation is the second step of protein synthesis. The mRNA interacts with cell machinery called ribosomes to create proteins. Blocks of transfer RNA (tRNA), each bound to an amino acid, is matched to its complementary mRNA strand. Then the amino acids separate from the tRNA and agglomerate to become a protein (Avila 1995).

Since mRNA is a precursor to protein synthesis, a biological assumption is often made that mRNA levels correspond to protein levels. Thus, biologists frequently use mRNA measurements as surrogate measurements of cell protein levels. New technologies developed over the last 20 years allow us to measure mRNA expression profiles with a higher degree of accuracy than ever before. Figure 1.4 provides an illustration of the transcription and translation process.

A gene expression microarray is a powerful method of measuring mRNA. A microarray chip is a slide where thousands of synthetic single strand DNA fragments are placed in a array configuration. Scientists then take the tissue sample in question and expose it to

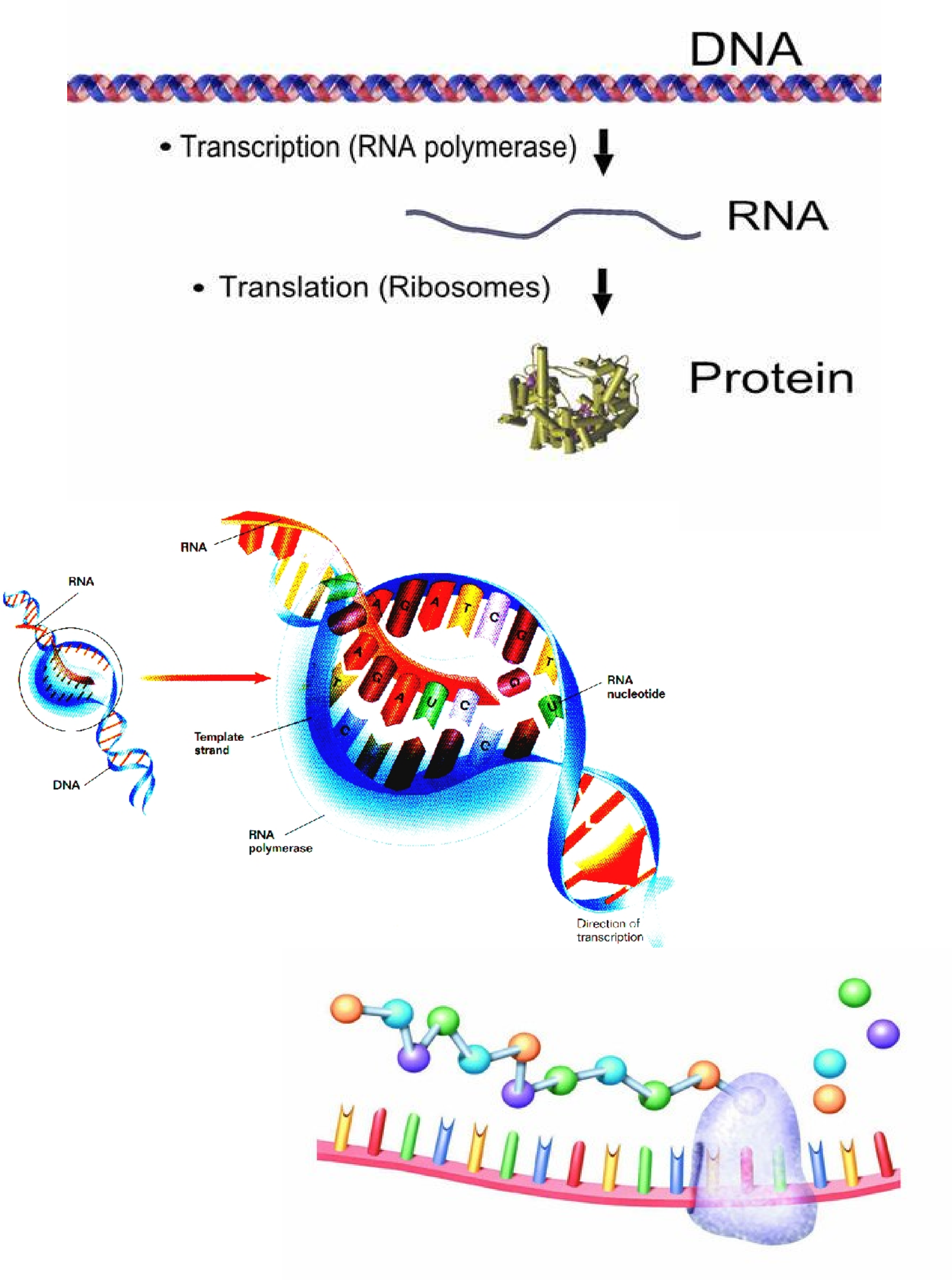

DNA
• Transcription (RNA polymerase)
RNA
• Translation (Ribosomes)
Protein
RNA
RNA
DNA
Template strand
RNA polymerase
RNA nucleotide
Direction of transcription

Figure 1.4: *Illustration of the transcription / translation process*

the microarray. When an mRNA sequence is complementary to a DNA sequence on the microarray, the two will hybridize (lock together). Then a fluorescent light in shone over the array, and the intensity at which the light is reflected is used as a measure of how much mRNA has hybridized to each sequence. The amount of mRNA that has hybridized to the sequence is then assumed to correspond to how much the gene in which the sequence is located is expressed for the organism cell-type (NIH 2012).

Modern technology such as DNA sequencing and microarrays allow us to quantitatively study the underlying relationship between differences in the genetic code, and observable physical traits. Understanding variation in gene expression is a central goal in medical and evolutionary genetics. Changes in the genetic code can result in biological conditions, affect

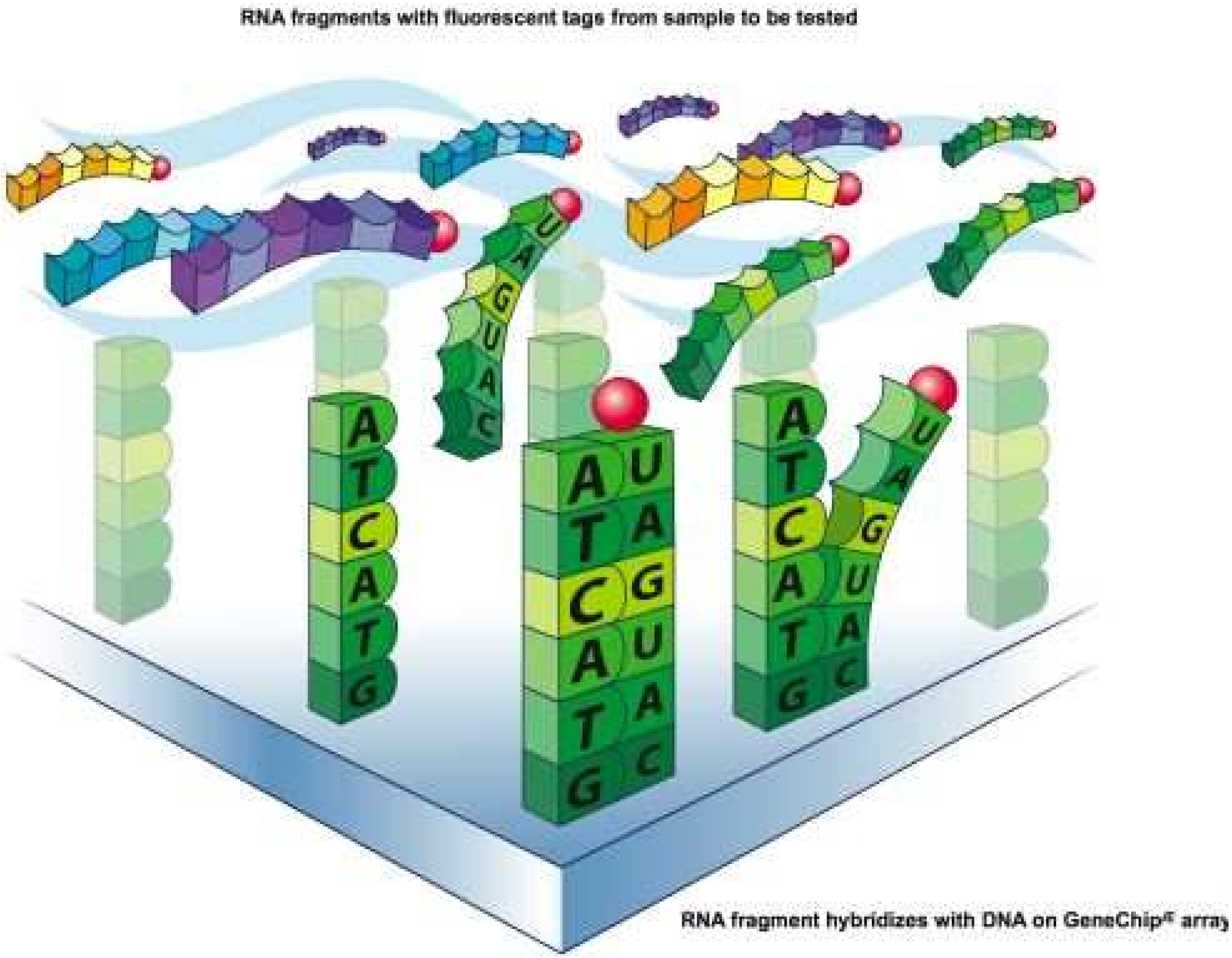

RNA fragments with fluorescent tags from sample to be tested
A T C A T G
A T C A T G
U A G U A C
A T C A T G
U A G U A C
U A G U A C
RNA fragment hybridizes with DNA on GeneChip® array

Figure 1.5: *Illustration of microarray slide*

how humans develop disease, and respond to pathogens, drugs, and vaccines. One can imagine a time when disease-causing mutations can be identified, and the underlying causal effects that result in the development of the disease can then be arrested through the use of drugs or other agents. As our scientific understanding of genetic variation increases, we become closer to realizing that goal.

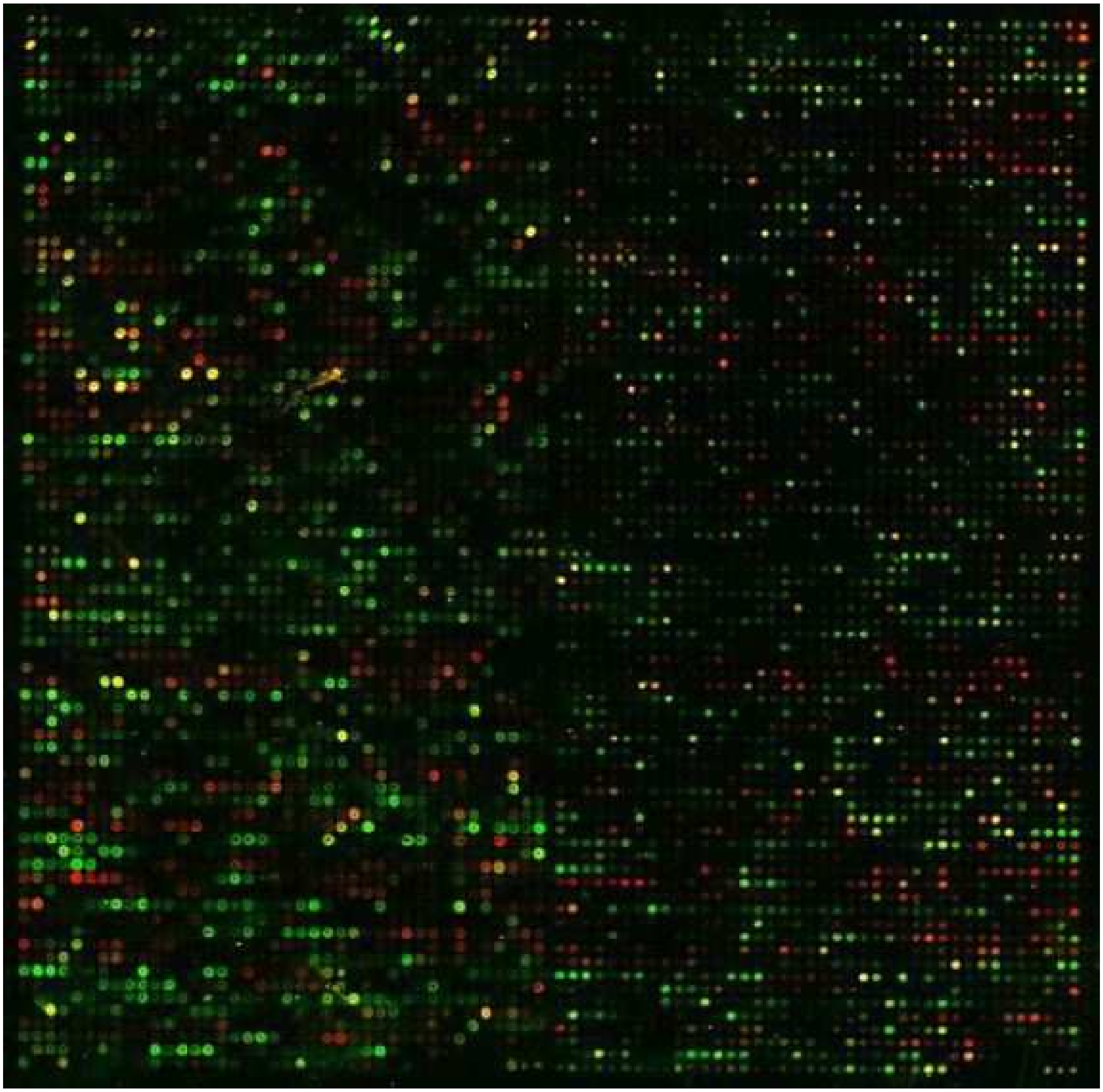


Figure 1.6: *Partial image of a lighted microarry chip*

1.2 Statistical model for eQTLs

Expression quantitative trait loci (eQTLs) are an important mechanism by which genetic variation occurs. An eQTL occurs when the expression level (amount of mRNA) of a gene is affected by a mutation in the genome of an individual. Where we have both genetic data and expression data for multiple individuals, we can use simple linear regression to model eQTLs. Figure 1.7 uses constructed data to illustrate this model. Visually, the figure seems to suggest that the number of minor alleles found at this particular SNP location has a nonzero linear relationship with the gene expression values as seen across a group of individuals. The probability of the linear association being nonzero can be tested using standard linear regression techniques.

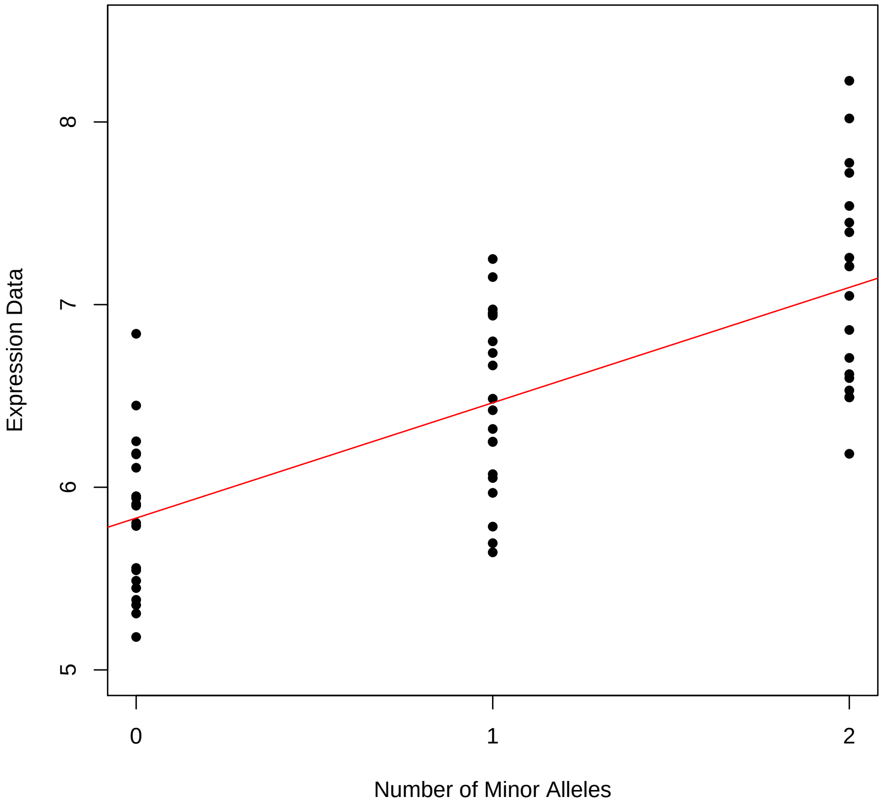


Figure 1.7: *eQTL example, using constructed data as an illustration.*

With the advent of DNA sequencing and procedures to measure gene expression, there have been many advancements in the field of genetics. One of the seminal eQTL studies was written by Lander and Botstein (1989). This paper showed how phenotypes of inbred organisms can be compared to SNPs found through restriction fragment length polymorphisms, to perform an extensive eQTL study. This was essentially a method of selective genotyping which used the EM algorithm to compute maximum likelihood estimations for the parameters of interest. A likelihood ratio test was then used to asses the probability of a nonzero association between SNP and phenotype in question.

Haley and Knott (1992) published a work which introduced the use of linear regression in place of maximum likelihood estimation to measure the relationship between SNPs and phenotype data. The paper emphasized the generalizeability of the model, and demonstrated that the model could be used to quickly and accurately estimate such relationships. Plomin and McClearn (1991) published one of the early eQTL studies of BXD mice (more on BXD mice later), assessing the relationships between genotypes and Mendelian traits.

Damerval *et al.* (1994) published one of the first papers which studied proteins from selected tissues as response variables for genotype data. This work used electrophoresis (sending an electric current through a fluid) to measure the quantity of anonymous proteins in a cross between two distinct lines of maize. Shena *et al.* (1995) wrote one of the first publications which used a microarray to measure gene expression. Cheung *et al.* (2004) compared the prevelance between *cis*-eQTLs (proximal) and *trans*-eQTLs (distant). In more recent years, the eQTL mapping studies of Veyrieras *et al.* (2007), and Pickrell *et al.* (2010), were a large source of motivation for this paper.

1.2.1 Regression Model

We present the simple linear regression model described by Veyrieras as the paradigm model for the relationship between gene expression and genotype. Suppose we have both SNP genotypes and expression measurements of a particular gene for each of M individuals.

Let y_j be the gene expression data for individual j (j in $1, \ldots, M$), and let g_{jk} be the genotype for individual j at SNP k (k in $1, \ldots, K$). Genotypes are coded as having 0, 1, or 2 copies of the minor allele. Then the effect of individual j's genotype at SNP k on his/her gene expression level is assumed to follow the additive linear model:

$$y_j = \mu + \beta_k \, g_{jk} + \epsilon_{jk} \tag{1.1}$$

where μ in the mean expression level for individuals with $g = 0$, and β_k is the additive effect of the minor alleles at SNP k, and ϵ_{jk} is the random error.

Table 1.1 shows a sample of what typical gene expression data and genotype data might look like. When we have multiple genes and SNPs we can extend model 1.1 as follows. Suppose we have expression data for M individuals over N genes, and genotype data for these M individuals over K SNPs. Let y_{ij} be the gene expression level of individual j (j in $1, \ldots, M$) at gene i (i in $1, \ldots, n$), and let g_{jk} be the genotype for individual j at SNP k (k in $1, \ldots, K$). Then we have:

$$y_{ij} = \mu_i + \beta_{ik} \, g_{jk} + \epsilon_{ijk} \tag{1.2}$$

where μ_i is the i^{th}-gene-level mean for $g = 0$, β_{ik} is the additive effect of the minor alleles of SNP k on gene i, and ϵ_{ijk} is the random error.

1.3 Heterogeneity

Gene expression data that is obtained from microarray technology is procured from a tissue sample placed on a microarray slide. These tissue samples are typically acquired the old-fashioned way, by a skilled technician with a scalpel. Each tissue sample is comprised of many cells. But these tissue samples are inherently comprised of multiple types of cells. So what is the effect of this heterogeneity of cell-types on observed expression profile of the tissue sample?

Probe	Gene	Chrome	Location	BXD1	BXD11	· · ·	BXD98	BXD99
1416734_at	Mkln1	6	173207123	7.290	6.796	· · ·	6.843	6.897
1416735_at	Asah1	8	59809599	11.165	10.792	· · ·	10.459	10.806
1416736_at	Casc3	11	72629825	10.543	10.816	· · ·	11.010	10.932
1416737_at	Gys1	7	60513192	9.250	9.420	· · ·	9.558	9.405
1416738_at	Brap	5	134623928	9.981	9.928	· · ·	10.259	10.187
1416739_a_at	Brap	5	89019109	7.352	7.964	· · ·	7.509	7.310
1416740_at	Col5a1	2	<NA>	6.866	6.923	· · ·	7.105	7.025
1416741_at	Col5a1	2	84605517	7.606	7.688	· · ·	7.567	7.701
1416742_at	Cfdp1	8	115916275	10.817	11.038	· · ·	10.816	11.359
1416743_at	<NA>	<NA>	119063096	9.174	8.896	· · ·	8.699	9.155
1416744_at	<NA>	<NA>	8845486	4.754	4.812	· · ·	4.736	4.841
1416745_x_at	<NA>	<NA>	17366695	5.085	4.971	· · ·	4.852	5.116
1416746_at	H2afx	9	135699737	10.246	10.323	· · ·	10.349	10.138
1416747_at	Mfsd5	15	103613440	10.347	10.552	· · ·	10.666	10.458
1416748_a_at	Mre11a	9	86474970	7.507	7.292	· · ·	7.484	6.903

SNP	Chrome	Location	BXD1	BXD11	· · ·	BXD98	BXD99
rs13482141	14	39961750	2	0	· · ·	2	0
gnf02.003.251	<NA>	<NA>	0	2	· · ·	0	2
rs13476330	2	5767413	0	2	· · ·	0	0
rs6300458	14	105565103	2	2	· · ·	2	0
rs13476286	1	187711002	2	2	· · ·	2	0
gnf18.027.000	<NA>	<NA>	0	0	· · ·	2	2
rs3712063	10	67552467	2	2	· · ·	2	2
CEL-6_17260153	<NA>	<NA>	2	0	· · ·	0	0
rs3665911	9	35793271	2	0	· · ·	2	2
rs13481201	11	103038835	0	2	· · ·	0	2
rs3700286	2	102471685	2	0	· · ·	2	0
rs3672808	6	139805730	0	2	· · ·	0	0
D4Mit178	4	66843132	0	2	· · ·	0	0
rs13479338	7	81203872	1	2	· · ·	2	2
rs13483716	20	10022276	2	2	· · ·	2	0

Table 1.1: *Samples of data from datasets GN110 and bxd.geno*

Suppose that our tissue sample is composed of two different cell-types, let us call these cell-types A and B. If we can assume that the ratio of mRNA fragments from each cell-type that hybridize with each probe set is the same as the ratio of cells in the tissue sample between cell-types, then microarray data would have the composition

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \gamma \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} + (1 - \gamma) \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \tag{1.3}$$

where y_i is the conglomerate amount that probe i is expressed, a_i is the amount which probe i is expressed in cell-type A, and b_i is the amount which probe i is expressed in cell-type B. γ_i is the proportion of cells from cell-type A of which the tissue sample is comprised.

From a general perspective, heterogeneity of cell-types in gene expression data present a number of issues that can affect downstream analyses. Microarray expression measurements from heterogeneous cell-types reflect weighted averages of expression levels within cell-type populations. When the cell-type expression profiles of the spurious cell-types differ from the cell-type of interest, this has the potential of providing misleading results regarding the expression profile of the cell-type of interest. Moreover, observed changes in gene expression may be the result of changes in the genetic code; or else they could be the result of changes in the abundance of an cell-type within the tissue. Conversely, genuine changes in expresion levels resulting from eQTL mechanisms may not be detected due to changes in the abundance of cell-types within the tissue.

1.3.1 Previous Studies of Heterogeneity

There have been a number of studies regarding the issue of heterogeneity of cell-types in gene expression data. An initial attempt was made by Venet *et al.* (2001), who proposed a linear model to estimate cell-type proportions and cell-type specific gene expression profiles. However, Venet relied on known prior information regarding the true gene expression profiles of the cell-types in question.

Lu *et al.* (2003) used a simulated annealing and knowledge of genes expressed only during specific cell cycles to estimate the proportion of cell-types from a mixed sample of yeast cells. Wang *et al.* (2006) performed an analogous study to estimate the proportions of cell types of the mammary gland using measurements taken over the course of murine mammary gland development and Ras-induced mammary tumorigenesis (artificially induced tumors). Again, prior knowedge of cell-type specific profiles was necessary for both of these approaches.

Another approach uses prior knowledge of cell-type proportions, as estimated by pathologists, to extract cell-type specific information. Stuart *et al.* (2004) used linear regression models to estimate the expected cell-type expression as the regression coefficient. Ghosh (2004) used a hierarchical mixture model to model the data. A combination of methods of moments procedures and the expectation-maximization (EM) algorithm were then used to estimate the model parameters. There are, however, questions about the availability and reliability of the estimates of cell-type proportions made by pathologists. Erkkilä *et al.* (2010) proposed a Markov Chain Monte Carlo sampler to alleviate some of the issues caused by potential inacuracies of such proportion estimates.

Several works have been published with the goal of solely estimating cell-type proportions. Gosink *et al.* (2007) used a ranking of gene expression levels as candidates to ‘separate’ cell-types. Clarke *et al.* (2010) refined Gosink’s method by using summary statistics of the dataset to make the method more robust to noise inherent in gene expression data. Leek and Storey (2007) proposed a surrogate variable analysis methodology for estimating cell-type proportions.

1.4 Contribution of The Thesis

Prior to this study, no methodology has been developed to both estimate cell-type proportions and cell-type specific expression profiles without a priori information. Our methodology can extract the cell-type proportions and cell-type specific expression profiles for a two-cell model without prior knowledge of either the cell-type proportions or the individual cell-type expression profiles.

The information provided by this methodology will benefit scientists in a number of ways. Researchers may be interested in the cell-type proportions of a tissue sample. For example it may be valuable to know the proportion of cancerous cells in a given tissue. In another scenario we may be interested in comparing expression profiles across varying cell-types. Accurate estimates of the individual cell-type expression profiles will rely on being able to isolate the cell-type of interest within expression data.

One of the sources of motivation for this research was supplied by the Genotype-Tissue Expression (GTEx) project. GTEx “aims to study human gene expression and regulation in multiple tissues, providing valuable insights into the mechanisms of gene regulation and, in the future, its disease-related perturbations. Genetic variation between individuals will be examined for correlation with differences in gene expression level to identify regions of the genome that influence whether and how much a gene is expressed” (National Institute of Health, 2012). The project is currently in the process of collecting genetic and expression-level data for various tissues from multiple individuals.

In order to effectively study differences in gene expression a thorough and complete mapping of eQTLs is essential. However, heterogeneity of cell-types within a tissue sample may result in innacurate expression profiles for the cell-type of interest. Even the most carefully planned and executed manual dissection will inevitably result in multiple cell types within the sample. While there are some techniques for obtaining a pure-cell type, they are generally costly and time-consuming making them impractical for a wide-ranging study such as the GTEx project where many samples are needed. Allowing the presence of a second (or more) cell type in expression data will presumably result in a loss of power and increase of false positives. Therefore a computational method for modeling heterogeneity found within expression data that would remove some of the detrimental effects of multiple cell types would be a useful resource for the researcher studying genetic variation.

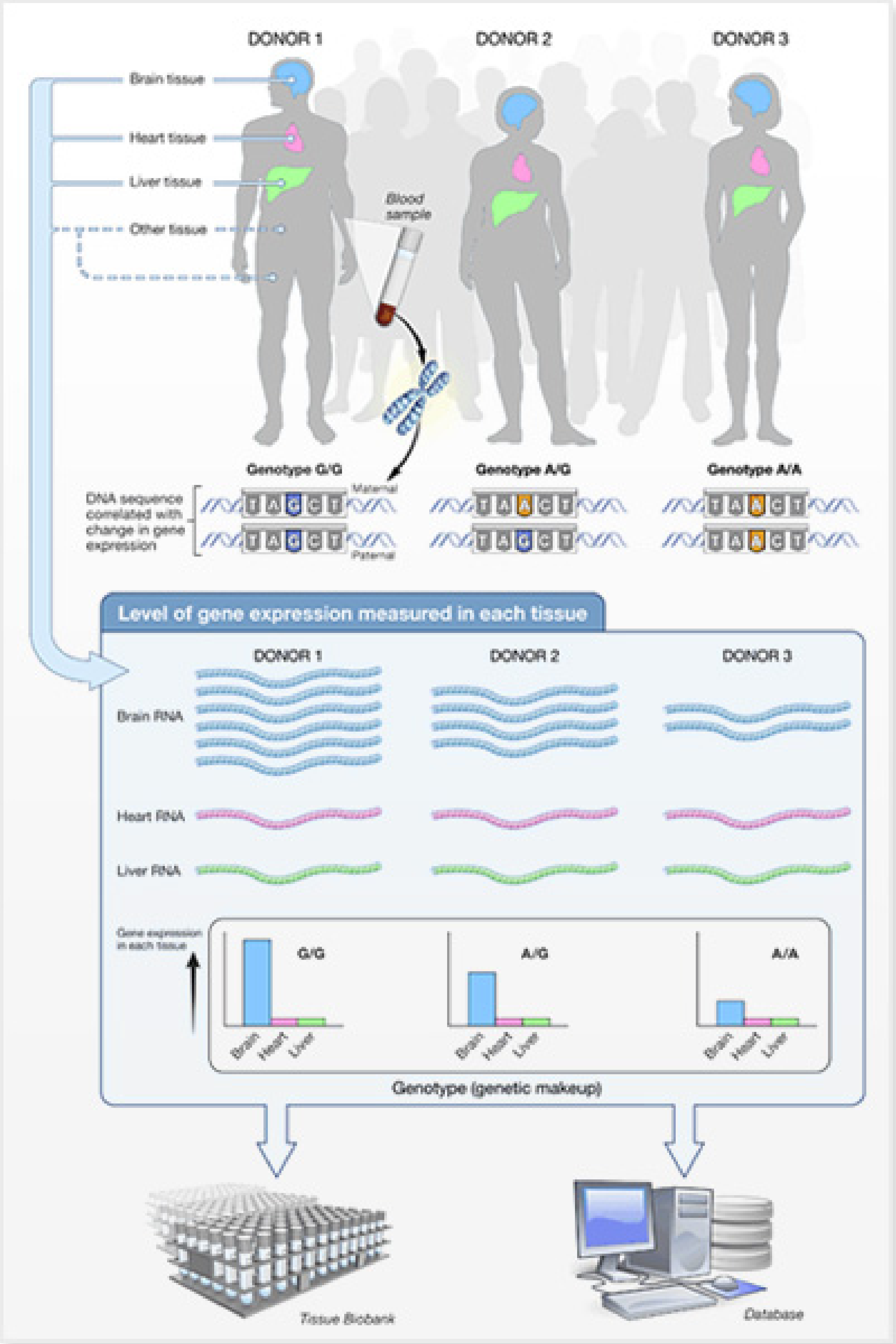


Figure 1.8: *The GTEx process*

1.5 Real Data Analysis

We use the parameter estimation methodology developed in this thesis to estimate the cell-type specific expression profiles and individual cell-type proportions for a real data set. We chose to analyze data collected from recombinant inbred (RI) strains of mice. Recombinant inbred strains of mice are mice that have been inbred for many generations until their genetic codes are virtually identical. These mice are then genotyped and bred with other strains. The progeny of the second generation of such breeding are then inbred until future generations of these offspring have identical genetic codes (but are unique across strains). Once these lines have been established, they can be reproduced and disseminated cheaply and easily.

Recombinant inbred mouse data has been used in many studies, and is often publicly available. Moreover, many of these datasets have undergone stringent quality control and error-checking processes (GeneNetwork 2012). These qualities make such datas ideal choices for our purposes. The largest and most widely studied strains of recombinant inbred mice are called the BXD strains. These mice are a mixture of two lines of mice; one called the 'B' line and one called the 'D' line.

The gene expression data used in our study, titled *GN110*, was downloaded from the GeneNetwork website [23]. The data was collected from the tissue samples using the Affymetrix 4.30 v2 chip. This chip uses a total of 45,101 probes. The samples were processed at the University of Memphis led by Thomas R. Sutter, with the actual processing steps performed by Shirlean Goodwin. The data is taken from hippocampus (region of the brain) tissue, in which we expect to have significant levels of heterogeneity. There are 205 mice in the dataset for a total dimension of $45,101 \times 205$. Out of the total of 205 mice, there were 138 mice with BXD genotypes, from which there were 69 distinct strains. The animals used to generate this set of data were obtained from the University of Tennessee, the University of Alabama-Birmingham, or directly from the Jackson Laboratory (original breeders of the BXD lines).

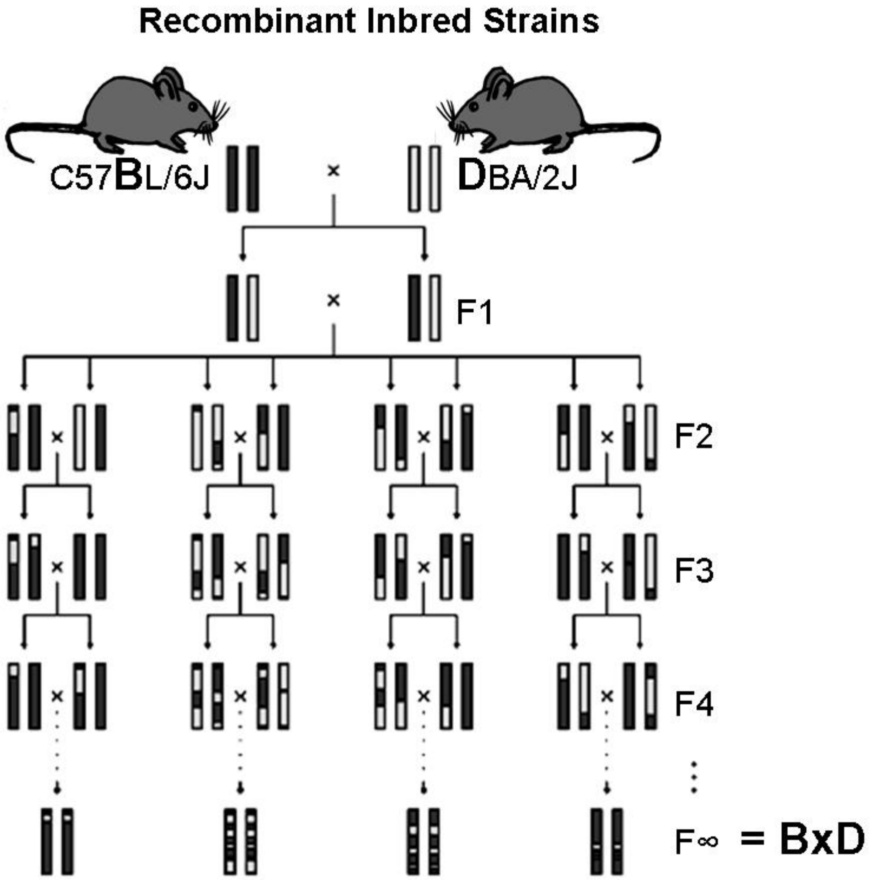


Figure 1.9: *Breeding process of BXD strain mice*

The BXD genotype data was also downloaded from the GeneNetwork website. The genotype file contains 3,796 markers over a total of 88 BXD strains. These locations have been determined by biologists to be places on the genome where the parental strains of the BXD mice have different alleles. The data for each recombinant inbred strain at each SNP is coded as either 'B,' 'D,' or 'H' for heterozygous. "SNPs are spaced approximately one every 4.3kb across the genome and were selected to be highly polymorphic among characterized mouse strains. Genotypes called from analysis of the array data are highly reliable. From an internal study of two strains, genotypes from 99.7% of the polymorphic SNPs that had genotypes in the NCBI dbSNP database had matching genotypes from the Diversity Array" (GeneNetwork, 2012).

Genotyping was described in Peirce *et al.* (2004). “Genotyping was performed the PCR protocol. DNA for the initial genotyping pass was purified from Princeton tail samples using standard phenol-chloroform extractions from single animals that contributed to the subsequent generations. Primer pairs purchased from Research Genetics (Huntsville, AL) were amplified using a high-stringency touchdown protocol in which the annealing temperature was lowered progressively from 60 °C to 50 °C in 2 °C steps over the first 6 cycles. After 30 cycles, PCR products were run on cooled 2.5% Metaphor agarose gels (FMC Inc., Rockland ME), stained with ethidium bromide, and photographed. Gel photographs were scored and directly entered into relational database files.”

The annotation data for the Affymetrix mouse 430 v2 chip was downloaded from the Bioconductor website. “Bioconductor is an open source, open development software project to provide tools for the analysis and comprehension of high-throughput genomic data. The Bioconductor project started in 2001 and is overseen by a core team, based primarily at the Fred Hutchinson Cancer Research Center, and by other members coming from US and international institutions” (Bioconductor, 2012).

1.5.1 Processing Data

The *GN110* dataset has dimensions $45,101 \times 205$ composed of 45,101 Affymetrix Mouse 4.30 v2 probes and 205 mice. We removed all mice that were not of a BXD genotype, which reduced the set to $45,101 \times 138$. Many of the probes had not been mapped to a location on the genome. We went through and removed all probes that did not have location annotation data. This reduced the total number of probes to 33,887. For this study we were interested in mapping the genes to local (nearby) SNPs. Many of the probes in the Affymetrix platform map to the same genes. When multiple probes mapped to the same gene, we took the mean expression levels of the probes across individuals to combine the data into gene level information. This further reduced the data to dimensions $17,155 \times 138$.

The genotype data also had some genotypes without location data. In addition, since the expression data did not contain probes from the X or Y chromosomes we did not use genotype data from these chromosomes. These entries were removed, resulting in dimensions $3,176 \times 88$. Table 1.10 and figure 1.11 provide some summary statistics of the processed dataset.

	Mean
Min.	4.434000
1st Qu.	6.961000
Median	8.143000
Mean	8.255000
3rd Qu.	9.380000
Max.	14.840000

Type	Frequency
B	174,497
H	6,981
D	171,550

Figure 1.10: *Summary statistics of the gene-level means / frequency table for genotype data*

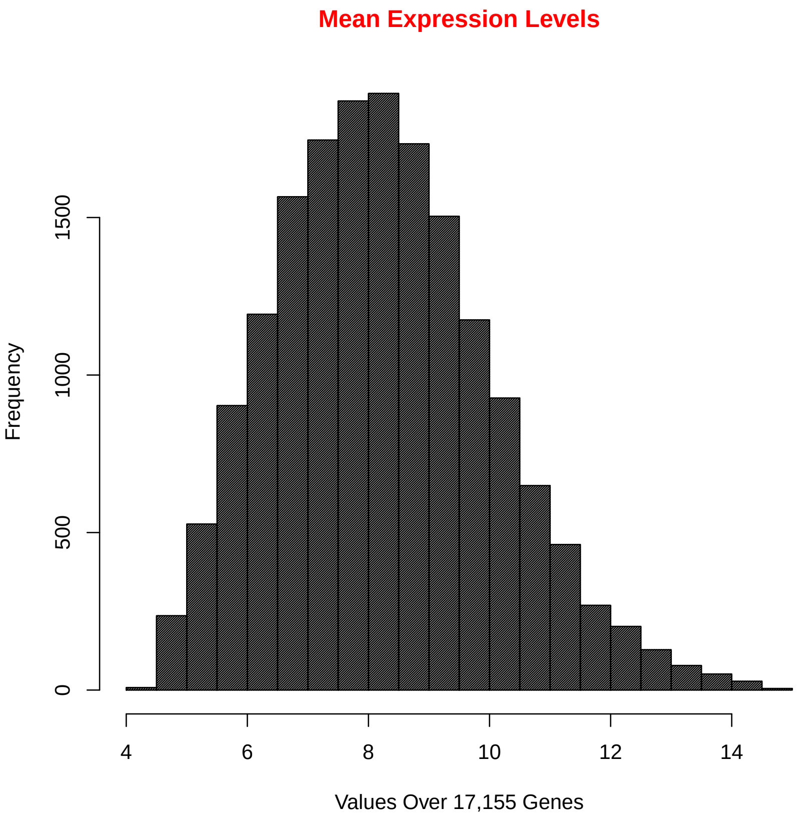


Figure 1.11: *Histogram of the gene-level means*

Chapter 2

Model

In developing the model used in this thesis, let us begin with a motivating example. Consider the following problem. Suppose we have hippocampus tissue samples from multiple individuals, and wish to estimate the expression profile of the hippocampus cells. However, it is well known that tissues in the brain are a heterogeneous mixture of neuron and glial cells. If we measure the gene expression levels of these tissues with a microarray, how can we estimate the true expression profiles of the hippocampus neuron and glial cells, and the cell-type proportions within tissue samples?

Our investigation begins with the linear regression model of the relationship between gene expression levels and genotypes previously described in Equation 1.2. Suppose we can describe the relationship between the neuron-cell expression profile and SNP genotype with the model

$$a_{ij} = \mu_i + \beta_{ik}\, g_{jk} + \epsilon^{(1)}_{ijk} \tag{2.1}$$

where each gene expression level, denoted a_{ij} is the sum of the gene-level mean plus the additive effect of the minor alleles (see Section 1.2 for a full description of this model). But for our hypothetical problem, we know that the observed expression data is the composition of two separate cell-types. Suppose that we can describe the relationship between the gene expression levels of glial-cells and genotypes of an individual at SNP k using the same model, but with different, unrelated, parameters values as follows

$$b_{ij} = \pi_i + \eta_{ik}\, g_{jk} + \epsilon^{(2)}_{ijk} \tag{2.2}$$

where b_{ij} is the glial-cell expression level for individual j at gene i, π_i is the expression level of gene i for $g = 0$, η_{ik} is the additive effect of the minor alleles for gene i and SNP k, g_{jk} is the genotype of individual j at SNP k, and $\epsilon_{ijk}^{(2)}$ is the random error.

Now recall the model for expression data comprised of two cell-types, previously described in Equation 1.3 (see Section 1.3 for more details).

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \gamma \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} + (1-\gamma) \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \tag{2.3}$$

This can be extended to multiple individuals by the following

$$\begin{aligned} \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1M} \\ y_{21} & y_{22} & \cdots & y_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{NM} \end{bmatrix} &= \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1M} \\ a_{21} & a_{22} & \cdots & a_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NM} \end{bmatrix} \begin{bmatrix} \gamma_1 & & & \Large 0 \\ & \gamma_2 & & \\ & & \ddots & \\ \Large 0 & & & \gamma_M \end{bmatrix} \\ &+ \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1M} \\ b_{21} & b_{22} & \cdots & b_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ b_{N1} & b_{N2} & \cdots & b_{NM} \end{bmatrix} \begin{bmatrix} 1-\gamma_1 & & & \Large 0 \\ & 1-\gamma_2 & & \\ & & \ddots & \\ \Large 0 & & & 1-\gamma_M \end{bmatrix} \end{aligned} \tag{2.4}$$

or in terms of a given y_{ij} we can express this as

$$y_{ij} = \gamma_j \, a_{ij} + (1-\gamma_j) \, b_{ij} \tag{2.5}$$

So we want to create a general model appropriate for a problem like the hypothetical one introduced at the beginning of this chapter. Suppose that we know the following conditions to be true:

- We have expression data which is comprised of exactly two cell-types
- The number of minor alleles has an additive effect on the expresion level of a gene for each cell-type
- The expression profiles of each cell-type are independent within individuals
- The proportion of the overall expression profile coming from each cell-type is independent across individuals

If these conditions are in fact met, then the full model which we use in this thesis is the composition of Equations 2.1, 2.2, and 2.5. Essentially, we have taken the equation for expression data comprised of two cell-types, and where we have expression levels a_{ij} and b_{ij} we have substituted the linear models for the individual expression-profiles of each cell-type.

We propose the following model as a two cell-type representation of the relationship between allele type and expression value for a given individual at a particular gene. Let us denote our model as follows:

$$y_{ij} = \gamma_j\,(\mu_i + \beta_{ik}\,g_{jk}) + (1 - \gamma_j)(\pi_i + \eta_{ik}\,g_{jk}) + \epsilon_{ijk} \tag{2.6}$$

where subscript i represents the ith gene, and j represents the jth individual. Let us denote cell-type A as the cell-type of interest, and cell-type B as the secondary cell-type. Then the parameters have the following meanings:

y_{ij} The gene expression value of individual j at gene i for a given tissue sample

γ_j The proportion of the tissue sample expression profile coming from cell-type A

μ_i The i^{th}-gene-level mean for $g = 0$, cell-type A

β_{ik} The additive effect of the minor alleles at SNP k on gene i for cell type A

g_{jk} The genotype of individual j at SNP k

π_i The i^{th}-gene-level mean for $g = 0$, cell-type B

η_{ik} The additive effect of the minor alleles at SNP k on gene i for cell type B

ϵ_{ijk} The random errors of the model

2.1 Objective Function Q

We use a least-squares approach to parameter estimation. Thus, we want to minimize the sum of the squared errors of our model with respect to μ, β, π, η, and γ. We begin with our model:

$$y_{ij} \quad = \quad \gamma_j\,(\mu_i + \beta_{ik}\,g_{jk}) + (1-\gamma_j)(\pi_i + \eta_{ik}\,g_{jk}) + \epsilon_{ijk} \tag{2.7}$$

Solving for ϵ and squaring both sides gives us

$$\epsilon_{ijk}^2 \quad = \quad \{y_{ij} - \gamma_j(\mu_i + \beta_{ik}\,g_{jk}) - (1-\gamma_j)(\pi_i + \eta_{ik}\,g_{jk})\}^2 \tag{2.8}$$

Let us define the objective function Q as

$$Q \;:=\; \sum_i \sum_j \epsilon_{ijk}^2 \;=\; \big\{y_{ij} - \gamma_j(\mu_i + \beta_{ik}\,g_{jk}) - (1-\gamma_j)(\pi_i + \eta_{ik}\,g_{jk})\big\}^2 \tag{2.9}$$

2.2 Alternating-Regression Algorithm

We utilize an alternating-regression method to estimate the parameters previously described in equation for the full model (equation 2.6). The alternating-regression method is a multi-step process that allows us to treat this equation as a linear model by fixing various parameters at different steps along the process. Being able to treat our model as a linear model at the various steps allows us to utilize standard linear regression theory to efficiently estimate the parameters at each step along the process. In Section 2.2 we demonstrate that fixing either γ or fixing μ, β, π, and η reduces the full model to a linear regression model. So these are the two sets of parameters which we alternatively fix as we minimize the objective function Q.

The alternating regression algorithm begins by fixing γ. γ was chosen as the starting point because we were able to develop a method of obtaining initial estimates for γ (described in section 2.2.4). Once the initial estimates for γ are obtained, then we consider these values to be fixed, and minimize the objective function Q, the squared loss of the fitted model, with respect to μ, β, π, and η.

Once we have estimates for μ, β, π, and η, we consider these values to be fixed, and minimize Q with respect to γ. These two minimization steps are repeated alternatively, until some convergence criterion has been achieved.

2.2.1 Fix γ and Minimize With Respect to μ, β, π, η

Recall the alternating regression algorithm. The recurring parameter estimation loop begins by fixing the values of γ and estimating values for μ, β, π, and η. Let us fix γ; then following least-squares approach, we want the values for μ, β, π, and η that will minimize the objective function Q. Thus, we have

$$\min_{\mu,\beta,\pi,\eta} Q = \min_{\mu,\beta,\pi,\eta} \sum_i \sum_j \left\{ y_{ij} - \gamma_j(\mu_i + \beta_{ik}\, g_{jk}) - (1 - \gamma_j)(\pi_i + \eta_{ik}\, g_{jk}) \right\}^2 \tag{2.10}$$

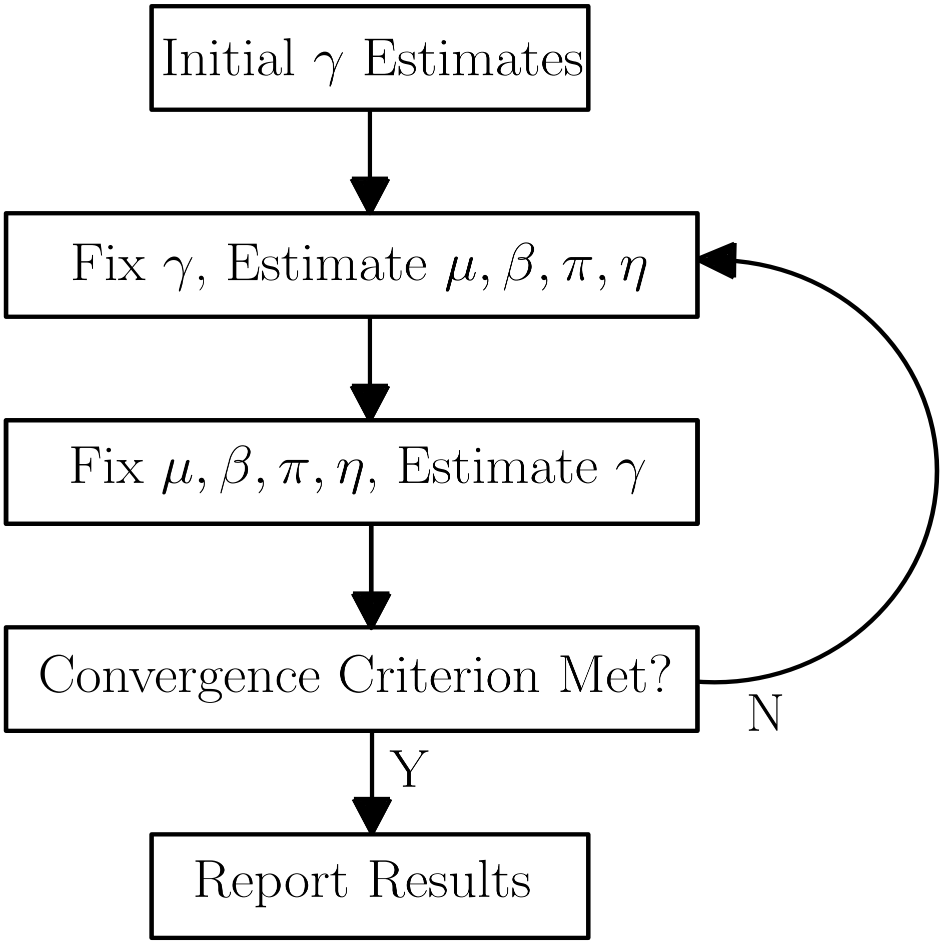

Initial γ Estimates
Fix γ, Estimate μ, β, π, η
Fix μ, β, π, η, Estimate γ
Convergence Criterion Met?
N
Y
Report Results

Figure 2.1: *Alternating-regression flow chart.*

$$= \sum_i \min_{\mu_i,\beta_i,\pi_i,\eta_i} \sum_j \left\{y_{ij} - \gamma_j(\mu_i + \beta_{ik}\, g_{jk}) - (1-\gamma_j)(\pi_i + \eta_{ik}\, g_{jk})\right\}^2 \tag{2.11}$$

Let us denote Q_i as

$$Q_i \; := \; \sum_j \left\{y_{ij} - \gamma_j(\mu_i + \beta_{ik}\, g_{jk}) - (1-\gamma_j)(\pi_i + \eta_{ik}\, g_{jk})\right\}^2 \tag{2.12}$$

Note that we can rewrite Q_i as

$$Q_i \;=\; \sum_j \{y_{ij} - \gamma_j\,\mu_i - \gamma_j\,g_{jk}\,\beta_{ik} + (1-\gamma_j)\,\pi_i + [\,(1-\gamma_j)\,g_{jk}]\,\eta_{ik}\}^2 \tag{2.13}$$

Then we can minimize each Q_i individually using a least squares linear model for

$$\min_{\mu_i,\beta_i,\pi_i,\eta_i} ||\,Q_i\,||^2 = \min_{\mu_i,\beta_i,\pi_i,\eta_i} ||\,Y - X\hat{\beta}\,||^2 \tag{2.14}$$

where the normed space is the Euclidean norm and we define Y, X, and $\hat{\beta}$ as

$$Y_{M\times 1} = \begin{bmatrix} | \\ y_{i\cdot} \\ | \end{bmatrix}, \quad X_{M\times 4} = \begin{bmatrix} | & | & | & | \\ \gamma & \text{diag}(\gamma)g_{\cdot k} & \text{diag}(1_M-\gamma)\pi & \text{diag}(1_M-\gamma)g_{\cdot k} \\ | & | & | & | \end{bmatrix},$$

$$\beta_{4\times 1} = [\;\mu_i \;\; \beta_i \;\; \pi_i \;\; \eta_i\;]' \tag{2.15}$$

If we can assume that $\epsilon_{i\cdot k} \sim N_M(0, \sigma^2 I_M)$, then $\hat{\beta} = (X^T X)^{-1} X^T Y$ is BLUE for β.

2.2.2 Fix μ, β, π, η and Minimize With Respect to γ

Recall the alternating regression algorithm. We treat our updated estimates of μ, β, π and η as fixed values. Now we use the least-squares approach to minimize γ. We want the values for γ that minimize the objective function Q, thus we have

$$\min_\gamma Q \;=\; \min_\gamma \sum_i \sum_j \left\{y_{ij} - \gamma_j(\mu_i + \beta_{ik}\,g_{jk}) - (1-\gamma_j)(\pi_i + \eta_{ik}\,g_{jk})\right\}^2 \tag{2.16}$$

$$= \min_\gamma \sum_j \sum_i \left\{y_{ij} - \gamma_j(\mu_i + \beta_{ik}\,g_{jk}) - (1-\gamma_j)(\pi_i + \eta_{ik}\,g_{jk})\right\}^2 \tag{2.17}$$

$$= \sum_j \min_{\gamma_j} \sum_i \left\{y_{ij} - \gamma_j(\mu_i + \beta_{ik}\,g_{jk}) - (1-\gamma_j)(\pi_i + \eta_{ik}\,g_{jk})\right\}^2 \tag{2.18}$$

Let us denote Q_j as

$$Q_j = \sum_i \left\{ y_{ij} - \gamma_j(\mu_i + \beta_{ik}\, g_{jk}) - (1-\gamma_j)(\pi_i + \eta_{ik}\, g_{jk}) \right\}^2 \tag{2.19}$$

Note that we can rewrite Q_j as

$$Q_j = \sum_i \left\{ (y_{ij} - \pi_i - \eta_i\, g_{jk}) - [\mu_i - \pi_i + (\beta_i - \eta_i)\, g_{jk}]\, \gamma_j \right\}^2 \tag{2.20}$$

Then we can minimize each Q_j individually using a least-squares linear model for

$$\min_{\gamma_j} ||\, Q_j\, ||^2 = \min_{\gamma_j} ||\, Y - X\hat{\gamma}_j\, ||^2 \tag{2.21}$$

where we define Y and X as

$$Y_{N\times 1} = \begin{bmatrix} | \\ y_{\cdot j} - \pi - \operatorname{diag}(\eta)\, g_{j\cdot} \\ | \end{bmatrix}, \quad X_{N\times 1} = \begin{bmatrix} | \\ \mu - \pi + \operatorname{diag}(\beta - \eta)\, g_{j\cdot} \\ | \end{bmatrix}, \tag{2.22}$$

and assuming $\epsilon_{\cdot jk} \sim N_N(0, \sigma^2 I_N)$, then $\hat{\gamma}_j = (X^T X)^{-1} X^T Y$ is BLUE for γ_j.

Since γ_j is a proportion, it has the constraint $0 \leq \gamma_j \leq 1$. So after estimation we perform the function $f(\gamma_j)$, where

$$f(\gamma_j) = \begin{cases} 0, & \gamma_j < 0 \\ \gamma_j, & 0 \leq \gamma_j \leq 1 \\ 1, & \gamma_j > 1 \end{cases} \tag{2.23}$$

2.2.3 Convergence Criterion

Convergence of estimates. Consider the objective function Q. At each step of the alternating regression process we are minimizing Q; thus we have a nonincreasing sequence of values for Q. Because Q is bounded from below by 0, there exists a minimum value for Q. Therefore, if we let the number of iterations of the alternating regression algorithm approach infinity, the value of Q will reach its minimum and the parameter estimates will have converged.

Since convergence of the parameter estimates of Q may require (potentially infinitely) many iterations, a convergence criterion at which to stop is needed. The criterion we set was as follows. Let $\hat{\mu}^{(k)}, \hat{\beta}^{(k)}, \hat{\pi}^{(k)}, \hat{\eta}^{(k)}$, and $\hat{\gamma}^{(k)}$ be the vectors of parameter estimates for the k^{th} iteration of the alternating-regression process. Then our convergence criterion was satisfied when (for $k \geq 2$)

$$\max \begin{Bmatrix} |\hat{\mu}^{(k)} - \hat{\mu}^{(k-1)}| \\ |\hat{\beta}^{(k)} - \hat{\beta}^{(k-1)}| \\ |\hat{\pi}^{(k)} - \hat{\pi}^{(k-1)}| \\ |\hat{\eta}^{(k)} - \hat{\eta}^{(k-1)}| \\ |\hat{\gamma}^{(k)} - \hat{\gamma}^{(k-1)}| \end{Bmatrix} < 0.001 \tag{2.24}$$

It should be noted that although we are guaranteed to achieve a local minimum for Q, there is no guarantee that this will be a global minimum, a common problem with optimization scenarios like this. Here we must rely on our initial estimates of the parameters. If our starting point is close enough to the global minimum with regards to the “smoothness” of the objective function, then we can avoid this problem.

2.2.4 Initial Estimates

Our algorithm requires that some initial estimates be provided. We assume that the greatest source of variation across individuals in our two cell-type model will come from differences in cell-type proportions. Thus, information about the cell-type proportions should be contained in the first principal components loading. After calculating the loadings, we then performed a location shift and scaled these loadings from the first PC to fit values between zero and one. The formula used to do this was

$$\hat{\gamma}_i = \frac{PC1_i - \min(PC1)}{\max(PC1) - \min(PC1)} \tag{2.25}$$

Chapter 3

Simulation

In this chapter, we use simulated data to test the performance of the parameter estimation methodology. Since we know the true values of the parameters for our simulated datasets, we can measure exactly how accurate the parameter estimates are, something that cannot be done with a real data set. Simulating data also lets us perform parameter estimation on many datasets, which allows us to study the consistency of the estimates.

The simulated datasets were created using parameters from the full model described in Chapter 2. That is to say, that each parameter from the model was simulated individually, and then a formula like the one in equation 2.6 was used to construct the simulated expression data. This expression data was then fed into the alternating regression algorithm, which tried to estimate the true parameters. The distributions for the parameters were loosely based on values from the real hippocampus data used in this thesis, as described in the following.

3.1 Simulating Parameters

- Simulating γ_j

The γ parameter is used to represent the proportion of cells in a tissue sample that come from cell-type A. Since γ is a proportion it must have a support between zero and one, inclusive. Also, we expect more of the density of β to be closer to one than to zero, because the tissue samples we are modeling are taken with the goal of obtaining as much of this cell-type as possible. For these reasons the beta distribution is a natural distribution for γ. The parameters were set as $\alpha = 1,\ \beta = 0.5$, which yields the density curve shown in Figure 3.1.

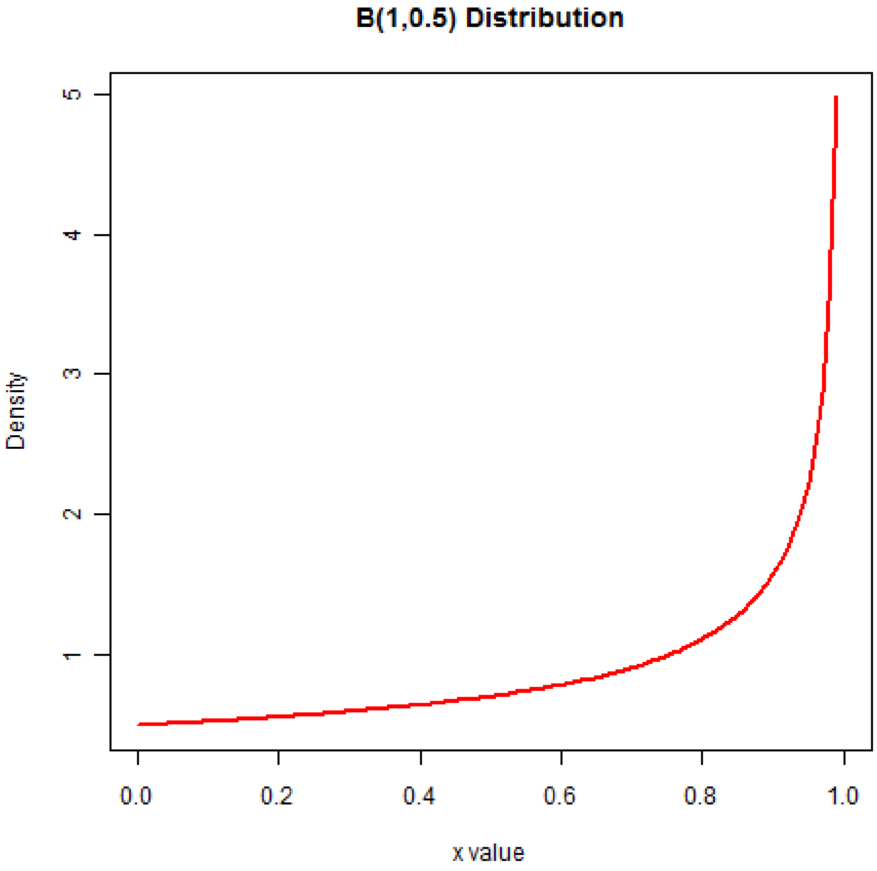


Figure 3.1: *Probability density function for a* $Beta(1, 0.5)$ *distribution*

- Simulating β, η

β and η are parameters which represent the additive effect of the minor alleles at a given SNP to the amount that a particular gene is expressed. We used the same process to simulate both β and η, so let us describe the simulation of β as an illustration of both. Consider two cases at an individual gene for $\beta_{ik}:$ one where there is an eQTL at SNP k for gene i, and one where there is no eQTL. We can model these cases with individually. Specifically, let

$$\beta_{ik} \overset{iid}{\sim} \begin{cases} N(\pm 1, {}^{1}\!/\!{}_{3}), & \text{SNP } k \text{ is an eQTL for gene } i \\ N(0, {}^{1}\!/\!{}_{10}), & \text{else} \end{cases} \tag{3.1}$$

Where 1/4 of the parameters were given the $N(\pm 1, 1/3)$ distribution. The values of ±1 were chosen arbitrarily as a convenient non-zero number. The η_i were then simulated using the exact same process.

- Simulating μ, π

μ and π are the gene-level means when the number of minor alleles is zero for cell-types A and B. The gene means are not assumed to have any particular distribution, so we gave them uniform distribution with values typical for the type of dataset used in this thesis. Specifically, let

$$\begin{aligned} \mu_i &\overset{iid}{\sim} \text{Unif}\,(4, 13) \\ \pi_i &\overset{iid}{\sim} \text{Unif}\,(4, 13) \end{aligned} \tag{3.2}$$

- Simulating Genotypes

The number of minor alleles for a particular SNP was determined using the following process. Two Bernoulli processes were simulated, representing the mother and the father of a diploid organism. The probability of success for each trial was set at 0.5 as a reasonable middle ground. Then the genotype was given by the number of successes for the two trials. This can be represented as the binomial random variable

$$g_{ij} \overset{iid}{\sim} \text{Bin}(\,2, 0.5) \tag{3.3}$$

Expression data matrix. Once all of the data was simulated, a perturbation factor was incorporated. We denoted $P_{N\times M}$ as our perturbation matrix, where each $p_{ij} \overset{iid}{\sim} N(0, \sigma^2)$. We used two different variances for our perturbation matrix; the 'small' variance was set as $\sigma^2 = 0.1$, and the 'large' variance was set as $\sigma^2 = 0.5$. Then the expression data matrix was constructed using the formula

$$y_{ij} = \gamma_j\,(\mu_i + \beta_{ik}\,g_{jk}) + (1 - \gamma_j)(\pi_i + \eta_{ik}\,g_{jk}) + p_{ij} \tag{3.4}$$

3.2 Parameter Estimation With Simulated Data

Simulation studies were performed on simulated data with two levels of random error introduced into the data. One set of simulations was run with the small error term added; these simulations were performed to measure how well the algorithm performs under ideal conditions. Then, a second set of simulations was performed with the larger error introduced. These simulations were run to get an impression of how well the parameter estimation method would perform under less than perfect conditions.

At each of these levels of error, we investigate parameter estimates for both a single simulation in detail, and then a meta-analysis of a run of 200 simulations.

3.2.1 Single-Simulation Results

In this section, the parameter estimates of a single data simulation at both levels of error are studied in detail. At various places statistics are presented regarding the parameter estimates for μ and β obtained using the standard linear regression model (equation 1.2) for comparitive purposes. In order to distinguish these estimates from the estimates obtained using the full model, these estimates are denoted as $\hat{\mu}^*$ and $\hat{\beta}^*$.

We begin our investigation of the two individual simulation studies with a look at the convergence rates of the simulations. Table 3.1 displays the maximum difference for each vector of parameters between consecutive iterations. The small perturbation and large perturbation simulations required 11 and 14 iterations to achieve convergence, respectively. These numbers are above average for both levels of error, as shown later in Table 3.3. As expected, the larger perturbation simulations tend to require more iterations to meet the convergence criterion.

Because the values for μ and π are on a larger scale than the other parameters, these tend to take the longest to pass the convergence threshold. In addition, it is the case that the observed expression data typically receives more of its weighted average from cell-type A (because of the distribution of γ). For this reason the values of π and η are harder to

	$\lvert\hat{\mu}_i^{(k)} - \hat{\mu}_i^{(k-1)}\rvert$	$\lvert\hat{\beta}_i^{(k)} - \hat{\beta}_i^{(k-1)}\rvert$	$\lvert\hat{\pi}_i^{(k)} - \hat{\pi}_i^{(k-1)}\rvert$	$\lvert\hat{\eta}_i^{(k)} - \hat{\eta}_i^{(k-1)}\rvert$	$\lvert\hat{\gamma}_i^{(k)} - \hat{\gamma}_i^{(k-1)}\rvert$
$k = 2$	0.010090	0.012041	0.017516	0.012726	0.000331
$k = 3$	0.000746	0.000497	0.002569	0.001350	0.000256
$k = 4$	0.000418	0.000201	0.002167	0.001066	0.000222
$k = 5$	0.000369	0.000172	0.001937	0.000938	0.000196
$k = 6$	0.000326	0.000152	0.001711	0.000828	0.000172
$k = 7$	0.000287	0.000134	0.001510	0.000731	0.000152
$k = 8$	0.000254	0.000118	0.001333	0.000645	0.000134
$k = 9$	0.000224	0.000104	0.001176	0.000569	0.000118
$k = 10$	0.000197	0.000092	0.001038	0.000502	0.000104
$k = 11$	0.000174	0.000081	0.000915	0.000443	0.000092

(a) Small perturbation

	$\lvert\hat{\mu}_i^{(k)} - \hat{\mu}_i^{(k-1)}\rvert$	$\lvert\hat{\beta}_i^{(k)} - \hat{\beta}_i^{(k-1)}\rvert$	$\lvert\hat{\pi}_i^{(k)} - \hat{\pi}_i^{(k-1)}\rvert$	$\lvert\hat{\eta}_i^{(k)} - \hat{\eta}_i^{(k-1)}$—	$\lvert\hat{\gamma}_i^{(k)} - \hat{\gamma}_i^{(k-1)}\rvert$
$k = 2$	0.056670	0.041972	0.060059	0.050186	0.000666
$k = 3$	0.003026	0.002482	0.004189	0.003034	0.000271
$k = 4$	0.000751	0.000349	0.002045	0.000902	0.000231
$k = 5$	0.000625	0.000283	0.001838	0.000787	0.000213
$k = 6$	0.000577	0.000264	0.001711	0.000729	0.000198
$k = 7$	0.000536	0.000245	0.001589	0.000677	0.000183
$k = 8$	0.000498	0.000228	0.001476	0.000629	0.000170
$k = 9$	0.000463	0.000211	0.001371	0.000584	0.000158
$k = 10$	0.000430	0.000196	0.001274	0.000543	0.000147
$k = 11$	0.000399	0.000182	0.001183	0.000504	0.000136
$k = 12$	0.000371	0.000169	0.001099	0.000468	0.000126
$k = 13$	0.000344	0.000157	0.001020	0.000435	0.000117
$k = 14$	0.000320	0.000146	0.000947	0.000404	0.000109

(b) Large perturbation

Table 3.1: *Covergence of parameter estimates between iterations*

estimate (consider the case when $\gamma = 1$), and as a result, usually π will be the last parameter to achieve convergence. Since small changes of β, η and γ can change our inferences regarding these parameters, we are happy to have this extra level of precision for them. Regardless, this structure of parameter convergence should be taken into account when setting an appropriate convergence criterion threshold.

Figure 3.2 displays scatterplots of estimates vs. actual values for small perturbations. All of the parameters are estimated extremely well under these ideal conditions; the R^2 value for each of these plots is at least 0.978. Even the formula to extract initial estimates for γ was able to provide near-perfect results.

The graphs for β and η do show some signs of innacurate estimates. The three distributions for the simulated values of β and η can be seen in these charts. In the case of η, it appears that there are some small problems with estimation when the true value of η is close to 0.

Figure 3.5, the scatterplots for the large perturbation, seems to develop this story even further. The R^2 values dip as low as 0.91 and 0.87 for β and η. The estimates are performing well as the true values of the parameters spread away from zero. But when the true estimates are near to zero, the estimation algorithm has some difficulty producing accurate estimates. When the true parameters are close to zero, but are in fact, nonzero the estimations are often being given as zero. In the context of eQTL mapping, this will result is type I errors. It does seem evident, however, that there are few true values close to zero being mistakenly estimated as nonzero.

Table 3.2 provides some more insight into the two simulations. Not surprising, the distributions of the absolute values of the errors are highly right-skewed. This is especially true for the paramaters of cell-type B: π and η. We can also compare the estimates μ and β for the full and simple regression models. As would be expected under these conditions, the full model outperforms the simple regression model in all categories.

Figures 3.4 and 3.5 provide a visual account of the summary statistics displayed in Table 3.2. Under ideal conditions, it can be seen that the full model is vastly outperforming the simple linear regression model. But it is interesting to note that with a larger random error factor, the median distances between estimates of β vs. actual values are rather comparable (0.11 vs. 0.22). However the upper bound for errors is much higher for the simple regression model than the full model (0.75 vs. 2.44). This is presumably caused by two main factors. Firstly, $^3/_4$ of the true values of β come from are normally distributed with mean 0. Since the median vaue for both estimation techniques will come from this distribution, we wouldn't expect the estimates to be too far off. The maximum errors for $\hat{\beta}^*$ are likely occuring when certain events transpire together. Consider, for example, the case when $\beta_{ik} = 1$, $\eta_{ik} = -1$, and $\gamma = 0.5I_M$. In essence, the additive effect of the minor alleles would cancel for both cell-types would cancel each other out, a very difficult scenario for the simple regression model to estimate.

	Median	Mean	Max.
$\lvert\hat{\mu} - \mu\rvert$	0.023896	0.028366	0.166494
$\lvert\hat{\mu}^* - \mu\rvert$	0.520157	0.613438	3.123300
$\lvert\hat{\pi} - \pi\rvert$	0.180963	0.208039	0.896013
$\lvert\hat{\beta} - \beta\rvert$	0.019482	0.023127	0.129806
$\lvert\hat{\beta}^* - \beta\rvert$	0.141075	0.192488	1.647710
$\lvert\hat{\eta} - \eta\rvert$	0.054444	0.067444	0.420750
$\lvert\hat{\gamma} - \gamma\rvert$	0.015116	0.026086	0.084588
$\lvert\hat{\gamma}_{\text{init}} - \gamma\rvert$	0.014283	0.026404	0.084588

	Cor.
$\text{Cor}(\hat{\mu}, \mu)$	0.999849
$\text{Cor}(\hat{\mu}^*, \mu)$	0.938161
$\text{Cor}(\hat{\pi}, \pi)$	0.994890
$\text{Cor}(\hat{\beta}, \beta)$	0.998538
$\text{Cor}(\hat{\beta}^*, \beta)$	0.870761
$\text{Cor}(\hat{\eta}, \eta)$	0.989167
$\text{Cor}(\hat{\gamma}, \gamma)$	1.000000
$\text{Cor}(\hat{\gamma}_{\text{init}}, \gamma)$	0.999977

(a) Small perturbation

	Median	Mean	Max.
$\lvert\hat{\mu} - \mu\rvert$	0.140213	0.168624	0.966488
$\lvert\hat{\mu}^* - \mu\rvert$	0.880538	1.021250	3.913900
$\lvert\hat{\pi} - \pi\rvert$	0.167807	0.200924	1.047130
$\lvert\hat{\beta} - \beta\rvert$	0.114517	0.137624	0.748990
$\lvert\hat{\beta}^* - \beta\rvert$	0.221030	0.301310	2.441560
$\lvert\hat{\eta} - \eta\rvert$	0.138122	0.164942	0.925134
$\lvert\hat{\gamma} - \gamma\rvert$	0.004515	0.004355	0.007375
$\lvert\hat{\gamma}_{\text{init}} - \gamma\rvert$	0.006781	0.006284	0.016525

	Cor.
$\text{Cor}(\hat{\mu}, \mu)$	0.994584
$\text{Cor}(\hat{\mu}^*, \mu)$	0.782634
$\text{Cor}(\hat{\pi}, \pi)$	0.992209
$\text{Cor}(\hat{\beta}, \beta)$	0.951996
$\text{Cor}(\hat{\beta}^*, \beta)$	0.680216
$\text{Cor}(\hat{\eta}, \eta)$	0.932811
$\text{Cor}(\hat{\gamma}, \gamma)$	0.999991
$\text{Cor}(\hat{\gamma}_{\text{init}}, \gamma)$	0.999833

(b) Large perturbation

Table 3.2: *Single simulation summary statistics*

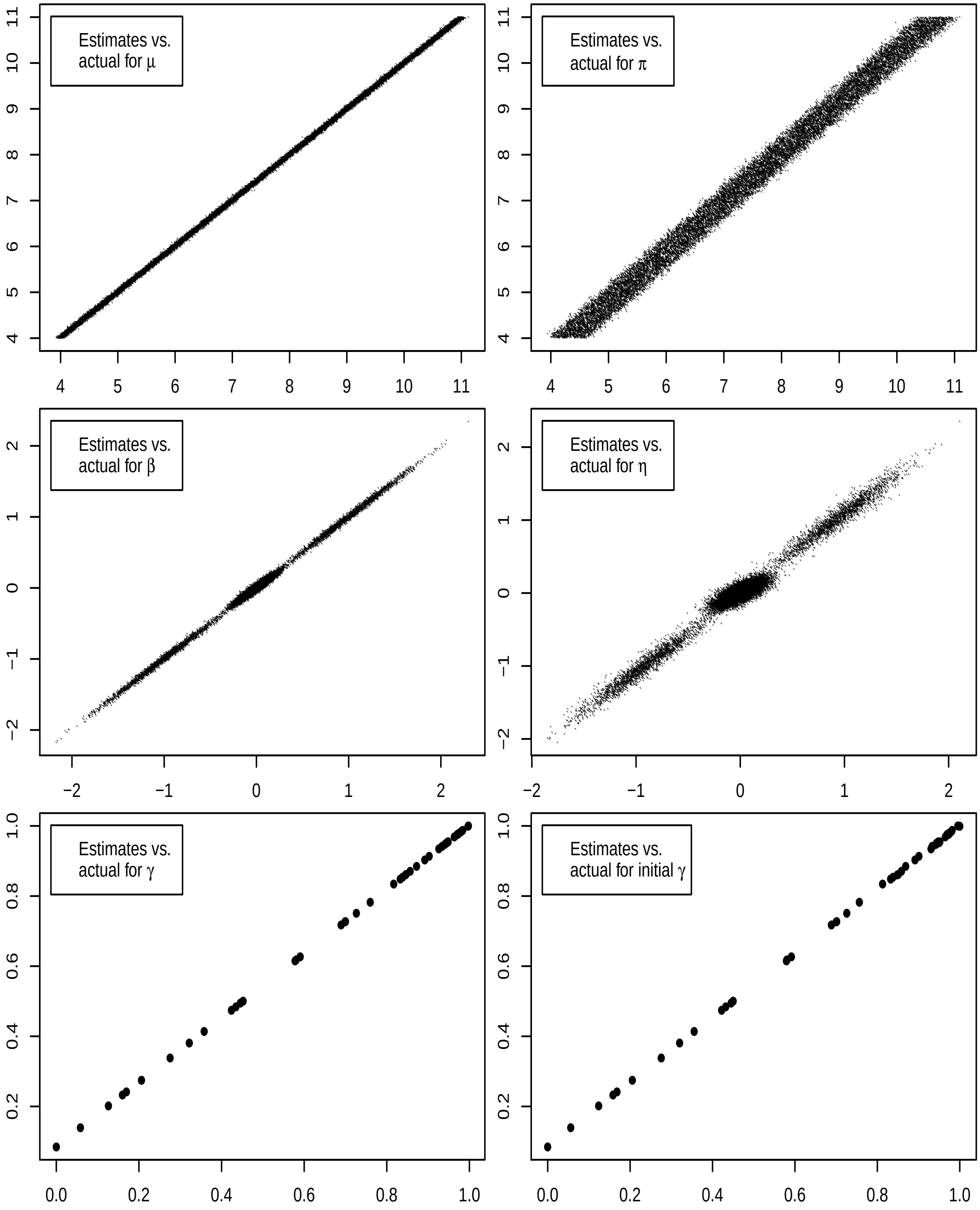


Figure 3.2: *Scatterplots of estimates vs actual values for small perturbation*

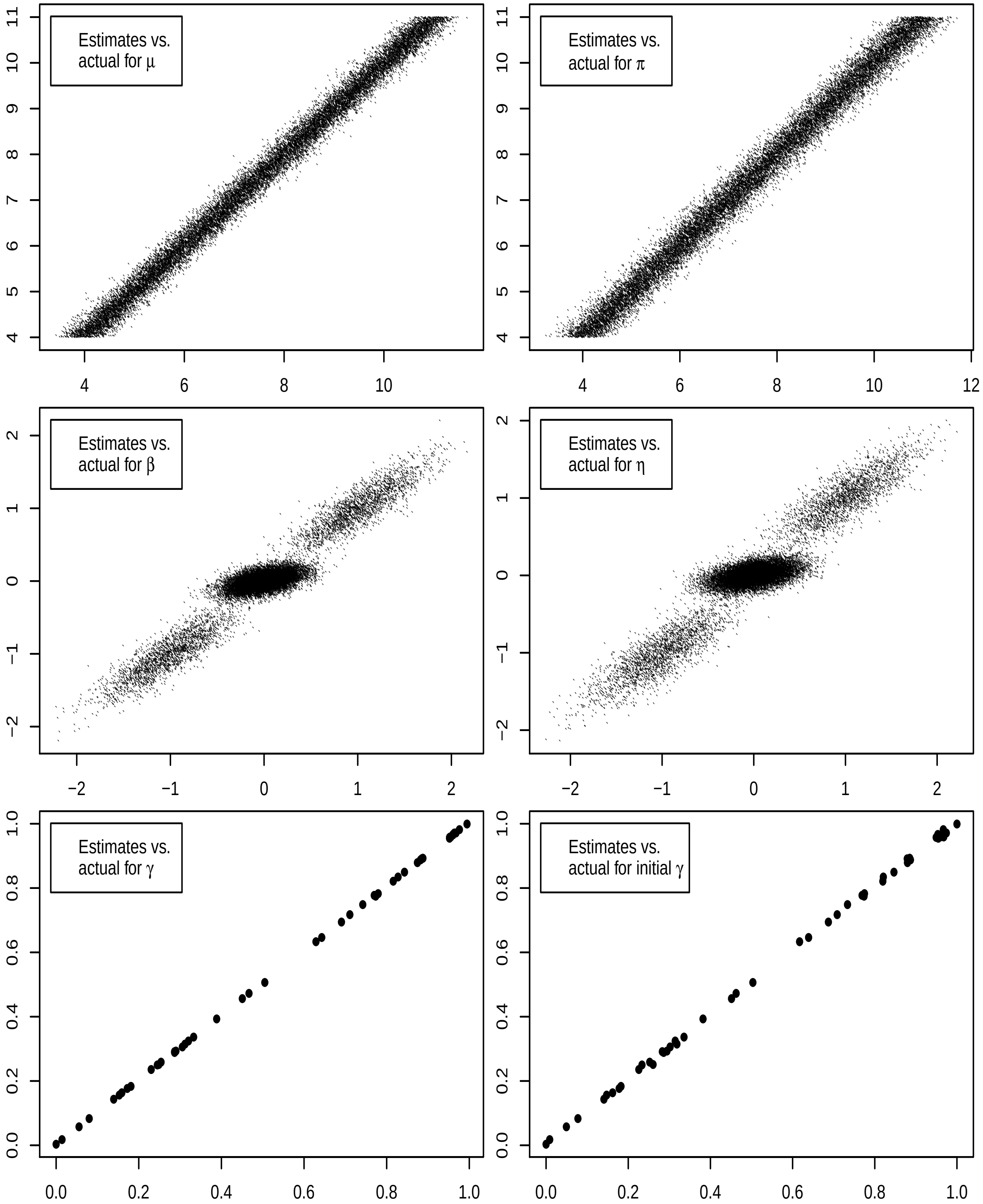


Figure 3.3: *Scatterplots of estimates vs actual values for large perturbation*

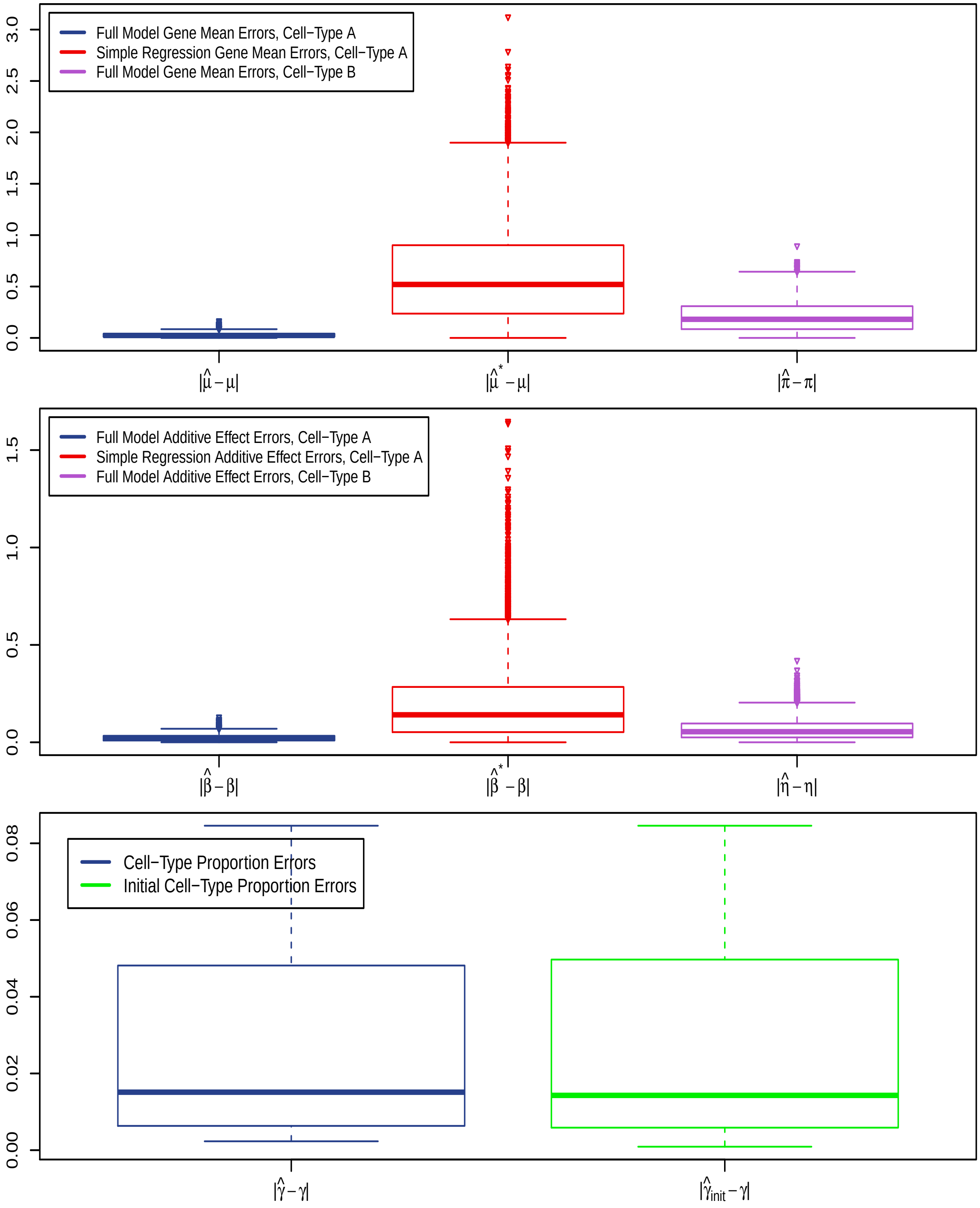


Figure 3.4: *Boxplots of the distances between estimates and actual values for small perturbation*

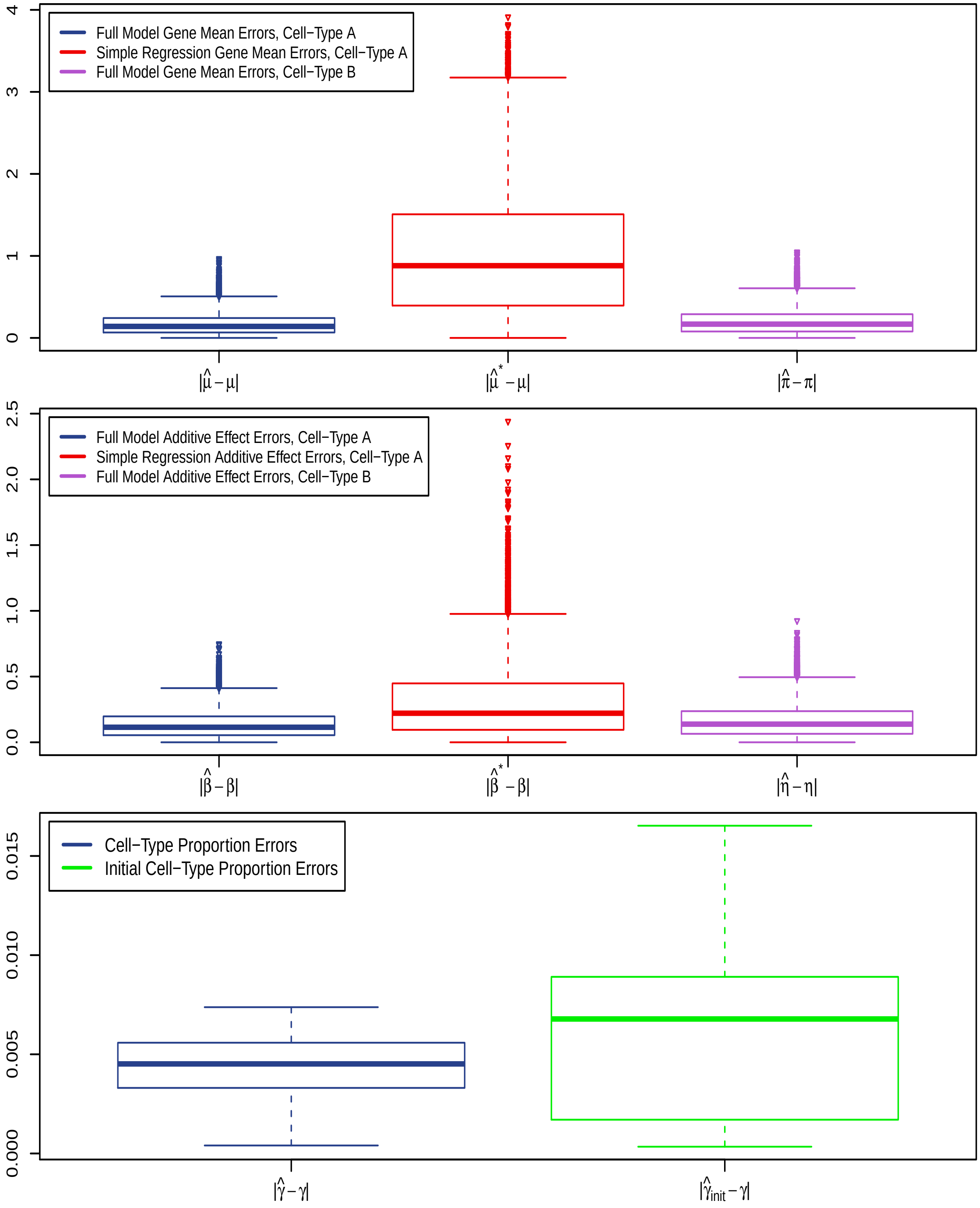


Figure 3.5: *Boxplots of the distances between estimates and actual values for large perturbation*

3.2.2 Multiple Simulations

In addition to the single-simulation studies that we did, we also did two runs of 200 simulations for both perturbation scenarios. Here we examine the findings from these simulations. Table 3.3 provides a summary of the number of iterations needed for each simulation to achieve convergence. It should be noted that over 50% of the simulations needed exactly four iterations to achieve convergence. Since the rate convergance determines the computing cost of estimating the parameters, it is encouraging that this number is reasonably low.

	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Small Perturbation	5	117	11	6	6	11	7	7	5	2	3	5	6	2	1	1	3	2	0	0	0
Large Perturbation	0	119	5	4	3	4	3	6	5	9	6	7	8	2	4	4	3	1	2	2	3

Table 3.3: *Number of iterations needed to achieve convergence*

Table 3.4 displays summary statistics for the data over the 200 iterations. The way the data for these tables was formed is as follows. The summary statistics (median, mean, maximum) of the absolute value of the differences between the parameter estimates and their true values was recorded for each of 200 simulations. The correlation coefficient between estimates and actual values was also recorded for each simulation. Next, the median and maximum entry out of the 200 simulations was recorded for each of those summary statistics. So, as an example, the $(1, 2)$ entry of Table 3.4 (a) would be the median value out of the 200 mean differences between $\hat{\beta}$ and β for small perturbation simulations.

The results from the multiple simulation studies seem to confirm the findings from the single simulation investigations. Under ideal conditions like the small perturbation simulations, the alternating regression algorithm was able to estimate parameters with a high degree of accuracy. The median correlation coefficient was at least 0.988 for all of the parameters estimated with the full model. The maximum single difference between the estimate of β and actual value for any simulation was 0.228.

Estimation was somewhat less accurate for the large perturbation simulations. Correlations coefficients for β and η were as low as 0.93 and 0.82, respectively (see figure 3.6). Correlation coefficients for γ remained greater than 0.999. The distribution of the errors is right-skewed, with a small percentage of errors very far from the medians. This is especially evident with larger perturbation, and for estimates of the cell-type B expression profile (π and η).

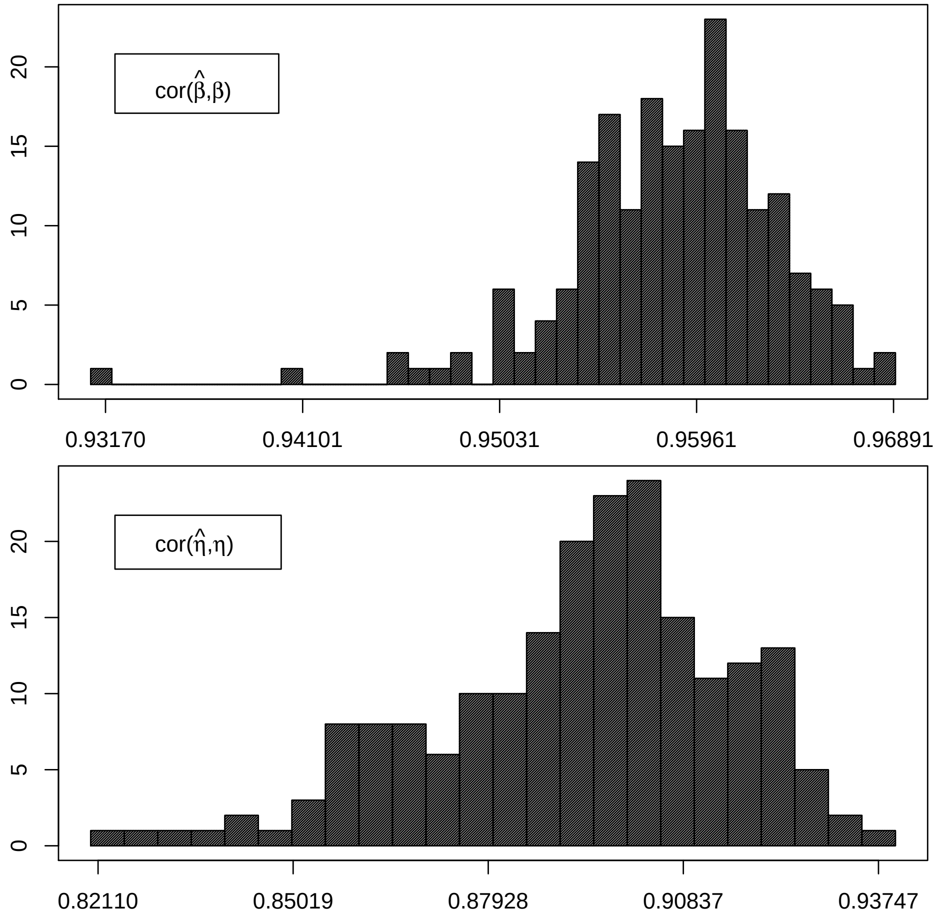


Figure 3.6: *Correlation coefficients between estimates and actual values for β and η over 200 simulations, large perburtation*

Figures 3.7 and Figure 3.8 provide boxplots of summary statistics for the parameters obtained over the course of the 200 simulations. Comparisons between errors for the parameter estimates of cell-type A for the full model and simple linear regression model are simular to those made for the single-simulations examination. The estimates made with the full model outperform those made with the simple regression model across the board. Again we see that the differences are especially demonstrable towards the right-tails or the error distributions. One interesting point to note is that the performance of the simple regression model does not suffer much between the small perturbation simulations and the large perturbation simulations (0.81 vs. 0.79 median correlation). One may conclude that the error resulting from underfitting the model is vastly greater than the error introduced from random error.

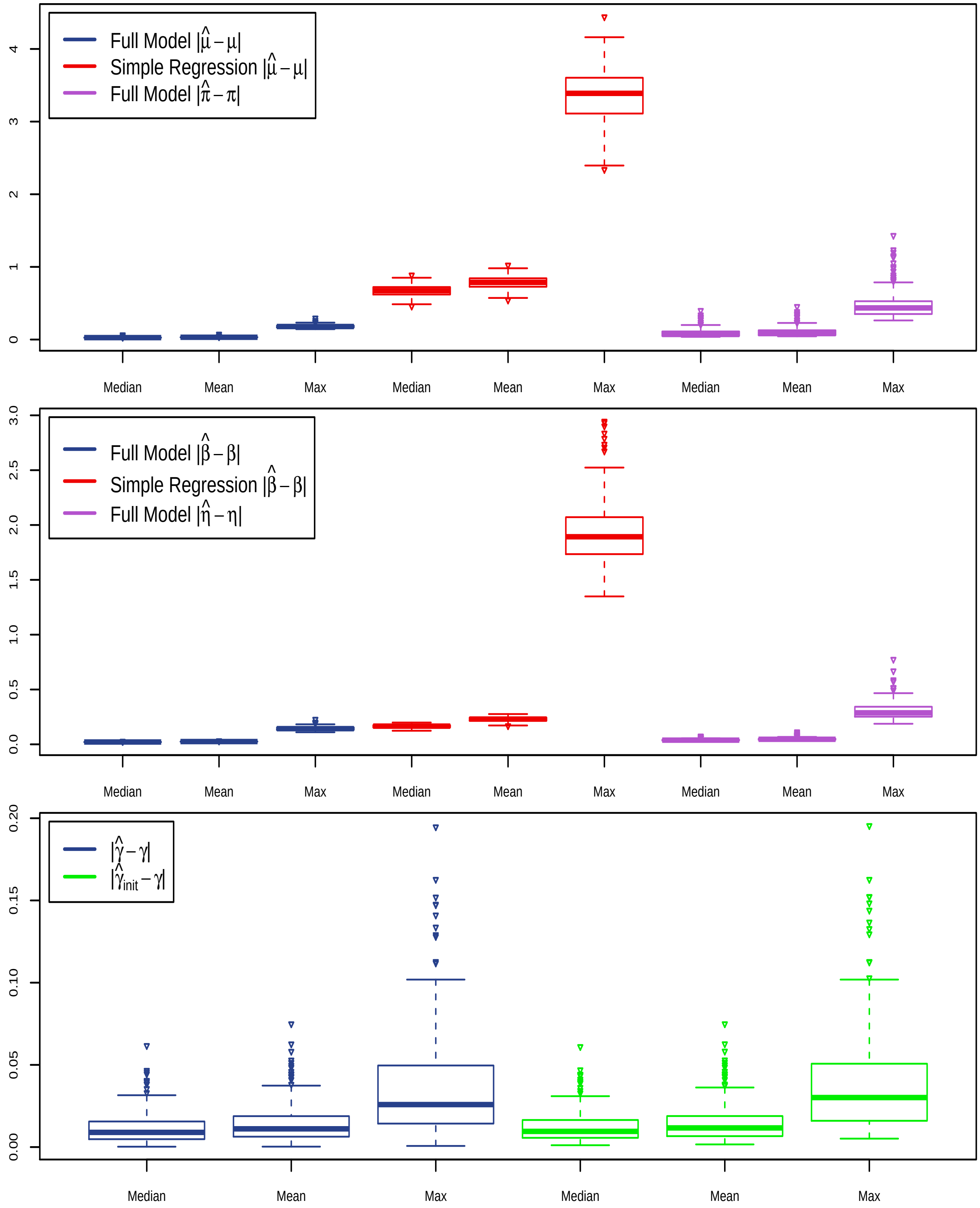


Figure 3.7: *Boxplots of the distances between estimates and actual values over 200 simulations, small perturbation*

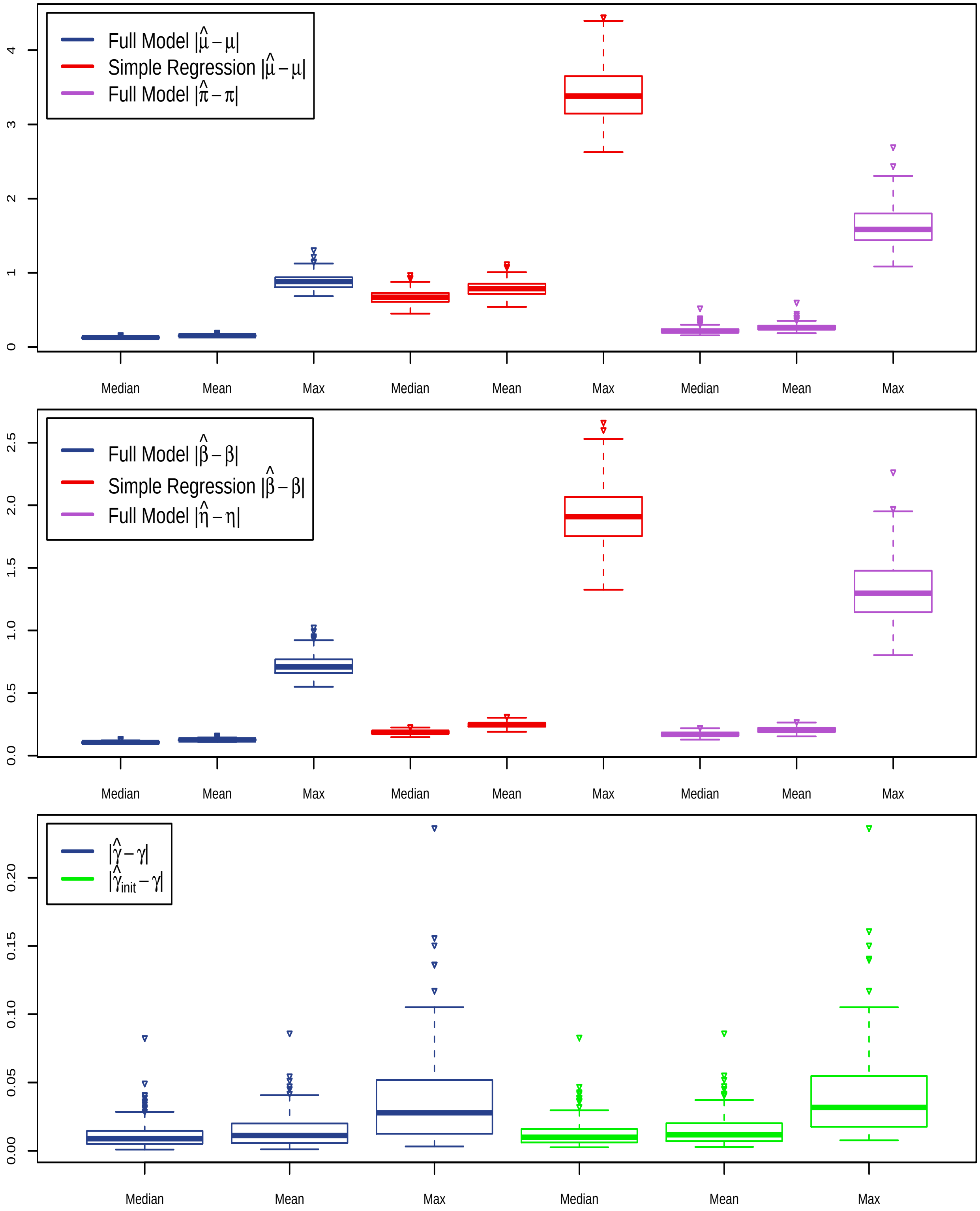


Figure 3.8: *Boxplots of the distances between estimates and actual values over 200 simulations, large perturbation*

			Summary statistics for the individual simulations			
			Median	Mean	Max.	Cor.
Summary statistics of the summary statistics	$\hat{\mu}$	median:	0.026187	0.031323	0.178983	0.999812
		max:	0.063421	0.072786	0.294510	0.999861
	$\hat{\mu}^*$	median:	0.672847	0.786515	3.389044	0.884439
		max:	0.886521	1.023250	4.440977	0.953637
	$\hat{\pi}$	median:	0.070221	0.081875	0.435217	0.999036
		max:	0.399066	0.453559	1.431716	0.999627
	$\hat{\beta}$	median:	0.021312	0.025382	0.142901	0.998233
		max:	0.026396	0.031377	0.227643	0.998688
	$\hat{\beta}^*$	median:	0.165979	0.230198	1.892718	0.807808
		max:	0.198904	0.276221	2.945403	0.907269
	$\hat{\eta}$	median:	0.038226	0.046128	0.287352	0.994336
		max:	0.073960	0.112940	0.774526	0.996869
	$\hat{\gamma}$	median:	0.008943	0.011099	0.025856	1.000000
		max:	0.061701	0.074904	0.194696	1.000000
	$\hat{\gamma}_{\text{init}}$	median:	0.009546	0.011660	0.030124	0.999966
		max:	0.061052	0.074925	0.195449	0.999985

(a) Small perturbation

			Summary statistics for the individual simulations			
			Median	Mean	Max.	Cor.
Summary statistics of the summary statistics	$\hat{\mu}$	median:	0.128135	0.152699	0.882586	0.995511
		max:	0.164744	0.197546	1.310403	0.996589
	$\hat{\mu}^*$	median:	0.670697	0.786063	3.383101	0.884690
		max:	0.974261	1.119224	4.448174	0.952732
	$\hat{\pi}$	median:	0.214646	0.257856	1.584960	0.986944
		max:	0.525513	0.603796	2.697766	0.993201
	$\hat{\beta}$	median:	0.105397	0.125379	0.708158	0.958917
		max:	0.136357	0.163212	1.025942	0.968915
	$\hat{\beta}^*$	median:	0.186571	0.247492	1.908560	0.789325
		max:	0.229187	0.313412	2.662000	0.881143
	$\hat{\eta}$	median:	0.168837	0.202277	1.297033	0.896312
		max:	0.223872	0.271100	2.265407	0.937465
	$\hat{\gamma}$	median:	0.008864	0.011187	0.027835	0.999991
		max:	0.082810	0.086217	0.236570	0.999995
	$\hat{\gamma}_{\text{init}}$	median:	0.009903	0.011746	0.031754	0.999918
		max:	0.083123	0.086240	0.236570	0.999958

(b) Large perturbation

Table 3.4: *Multiple simulations summary statistics*

Chapter 4

Real Data Analysis

In this chapter, the expression profile of mouse hippocampus data is estimated. The data used was BXD mouse data downloaded from the University of Tennessee GeneNetwork website (see section 1.5 for details). This dataset contains data for 138 mice with BXD genotypes; of these 138 mice there are 69 distinct BXD genotypes. It is standard procedure when expression data is measured to take measurements of replicate genotypes when possible. This allows biologists to take an average between expression profiles of mice with the same genotype, which reduces the variation from certain factors like measurement error and heterogeneity of cell-types in tissue samples. But since this heterogeneity is exactly what we wish to capture, we take a different approach.

Instead of taking an average between mice with the same genotype, we selected exactly one mouse per genotype and estimated the expression profiles for the two-cell model. Being able to choose from some number of mice for each genotype (there were either 1, 2, or 3 for each) allows us to select many different permutations of mice and perform parameter estimation for each permutation. In the following sections we show some results for a single permutation, and then a run of 50 different permutations.

4.1 One Permutation Estimations

In this section we present the results from estimating the parameters for a single random permutation of mice. That is to say, that among the 69 genotypes of BXD mice, we randomly selected exactly one mouse from each genotype. The process took 14 iterations to achieve a convergence criterion level of 0.01 (the same criterion as described in equation 2.24 but with a relaxed convergence level to account for the messy nature of expression data). The convergence rate of the parameter estimates are summarized in table 4.1.

Table 4.2 provides paramater estimates for a randomly chosen subset of 35 genes, with p-values shown for the simple regression estimate of β. The correlation coefficients between the full model and the simple regression model for μ and β are 0.995 and 0.773, respectively.

	$\lvert\hat{\mu}^{(k)} - \hat{\mu}^{(k-1)}\rvert$	$\lvert\hat{\beta}^{(k)} - \hat{\beta}^{(k-1)}\rvert$	$\lvert\hat{\pi}^{(k)} - \hat{\pi}^{(k-1)}\rvert$	$\lvert\hat{\eta}^{(k)} - \hat{\eta}^{(k-1)}\rvert$	$\lvert\hat{\gamma}^{(k)} - \hat{\gamma}^{(k-1)}\rvert$
$k = 2$	0.304679	0.235951	0.612022	0.576787	0.542642
$k = 3$	0.229603	0.134413	0.514765	0.365494	0.200853
$k = 4$	0.115091	0.088030	0.314559	0.164319	0.105522
$k = 5$	0.080841	0.060893	0.170371	0.089323	0.050350
$k = 6$	0.059704	0.045010	0.128324	0.068947	0.038062
$k = 7$	0.045188	0.033028	0.095707	0.050165	0.028459
$k = 8$	0.034840	0.025678	0.072561	0.038550	0.021749
$k = 9$	0.026192	0.018137	0.055291	0.028971	0.015623
$k = 10$	0.018928	0.015241	0.041611	0.022432	0.012873
$k = 11$	0.015705	0.011890	0.031065	0.015521	0.009414
$k = 12$	0.010380	0.007134	0.022448	0.011878	0.006390
$k = 13$	0.008307	0.005385	0.017092	0.009700	0.004602
$k = 14$	0.006474	0.004253	0.013255	0.006314	0.004556
$k = 15$	0.004845	0.003689	0.009209	0.004525	0.002153

Table 4.1: *Real data single permutation convergence rate summary statistics.*

4.2 Multiple Permutations Summary Statistics

In this section we estimated the parameters for the *GN110* dataset using 50 distinct permutations of mice. The mice used in each permutation were chosen at random from set of between one and three mice for each particular genotype. The estimates were obtained using both the full model and the simple linear regression model so as to make comparisons between the two. Table 4.3 describes the amount of iterations needed to achieve convergence. The number of iterations ranged from 5 to 23, while the bulk of the iterations were between 8 and 11.

Gene	$\hat{\mu}$	$\hat{\mu}^*$	$\hat{\beta}$	$\hat{\beta}^*$	$\hat{\pi}$	$\hat{\eta}$	$P(\hat{\beta}^* = 0)$
Col6a2	9.44203	9.33269	0.03952	0.01314	9.15977	-0.01110	0.55820
Aacs	8.59271	8.51992	-0.00382	-0.00855	8.41598	-0.01849	0.49909
Aph1a	9.64702	9.39925	0.00515	0.01655	9.05920	0.01217	0.48458
Tbc1d9b	9.54571	9.47111	0.03861	0.01535	9.37316	-0.03348	0.32322
Elovl4	9.74437	9.69782	0.09855	0.10473	9.62975	0.11351	< 0.00001
A830080D01Rik	7.70850	8.22762	-0.01643	-0.02555	8.87166	0.06074	0.45313
Tpd52l2	8.25827	8.37211	-0.01436	-0.02319	8.53619	-0.03295	0.07526
Ankdd1b	8.45212	8.76414	0.07177	-0.00241	9.22364	-0.11222	0.88719
Tmem44	7.61123	7.60790	0.01300	-0.02162	7.60152	-0.07056	0.32420
2810021J22Rik	7.51027	7.70623	-0.02302	-0.01934	7.96289	0.03028	0.37315
4930430F08Rik	7.34618	7.66075	0.01605	0.00656	8.19208	-0.05883	0.75885
Rab33a	10.09197	10.18526	-0.03022	0.01357	10.31016	0.08875	0.50454
Fga	5.96737	5.99564	0.08854	0.03156	6.03505	-0.05417	0.32280
Tmem202	5.07375	5.10802	0.00558	-0.00995	5.16497	-0.03670	0.34053
Ror2	6.32052	6.29600	-0.00024	-0.00363	6.25956	-0.00787	0.63166
Elmod3	9.05404	8.94393	0.03523	0.02656	8.79623	0.00243	0.28053
Hyls1	7.33759	7.05757	0.03902	0.06111	6.62269	0.11808	0.02057
Steap1	7.11857	7.43024	0.04274	0.01192	7.98622	-0.12010	0.85632
Hlx	6.79844	6.68739	0.01100	-0.02529	6.51649	-0.07047	0.14869
ORF63	6.30348	6.21917	-0.03783	-0.00368	6.08208	0.05431	0.82314
Txnrd1	7.81803	7.77906	-0.03448	-0.04354	7.71439	-0.05029	0.00138
Synj2bp	9.72657	9.90656	0.13605	-0.00350	10.15370	-0.20705	0.83167
Spef2	5.37491	5.41132	0.00599	0.01385	5.47575	0.01608	0.23556
Iqcb1	7.56964	8.03789	0.09971	0.01584	8.78543	-0.14827	0.59665
Agpat4	11.23272	11.17328	-0.04110	-0.06435	11.08340	-0.09478	0.00021
Nptx2	7.17601	7.25361	0.09757	0.10909	7.38057	0.10550	< 0.00001
Jag2	8.63612	8.41494	0.05135	0.00645	8.03055	0.02097	0.78120
Dhx40	9.25713	9.43755	0.03804	0.02948	9.73506	-0.01006	0.10539
Pde6a	6.35925	6.29973	-0.01792	0.00558	6.20316	0.04589	0.59795
Lpar4	5.47823	5.74565	0.03303	-0.00184	6.12522	-0.04266	0.91482
Trip11	6.63798	6.73815	0.07852	0.18348	6.92785	0.27042	< 0.00001
Ttc30b	8.09382	8.45030	-0.08843	0.02829	8.99220	0.17370	0.43608
Nbr1	11.01091	10.99110	-0.03141	-0.00138	10.96388	0.04564	0.94283
Uqcrc1	13.08599	12.91979	-0.04256	-0.00555	12.68446	0.04236	0.74750
Rin2	8.77943	9.08552	0.02636	-0.06817	9.53553	-0.20738	0.01598

Table 4.2: *Real data single permutation: the parameter estimates from both the full model and simple linear regression model for a random subset of 35 genes.*

Table 4.4 displays summary statistics regarding the correlation coefficient between the full model estimates and the simple regression estimates of μ and β for each of the mouse permutations. Interestingly, there is quite a high degree of correlation between the estimates. Since the simple regression model is known to produce reasonably good results, some degree of correlation would indicate that the estimates from the full model are approximately where they should be.

Iterations	5	6	7	8	9	10	11	12	13	14	15	16	18	19	23
Frequency	1	4	3	8	5	5	8	3	3	2	2	1	2	2	1

Table 4.3: *Number of iterations required to achieve convergence*

	Min.	Median	Mean	Max.
$\text{cor}(\hat{\mu}, \hat{\mu}^*)$	0.995300	0.997100	0.997000	0.998300
$\text{cor}(\hat{\beta}, \hat{\beta}^*)$	0.768500	0.861400	0.851900	0.911300

Table 4.4: *Real data multiple permutations; summary statistics regarding the correlation coefficient between parameter estimates for the full and simple linear regression model.*

We also investigated the estimates of the cell-type proportion γ. Table 4.5 and Figure 4.1 both describe the median estimates of γ for each of the 138 mice. The frequency column in Table 4.5 refers to the number of times that a particular mouse was chosen to represent his or her genotype out of a possible 50 mouse permutations, ranging from 11 to 50 (50 occurs for those mice who were the sole representative of their genotype).

Figure 4.2 shows a histogram of correlation coefficients between the initial estimates and final estimates of γ for each of the 50 mouse permutations. The correlation coefficient for each of the 50 permutations ranges from 0.41050 to 0.78170 with a median coefficient of 0.62000. If our estimates for γ have any merit, then this would indicate that we are indeed picking out a good amount of information regarding the cell-type proportion from the PC loadings.

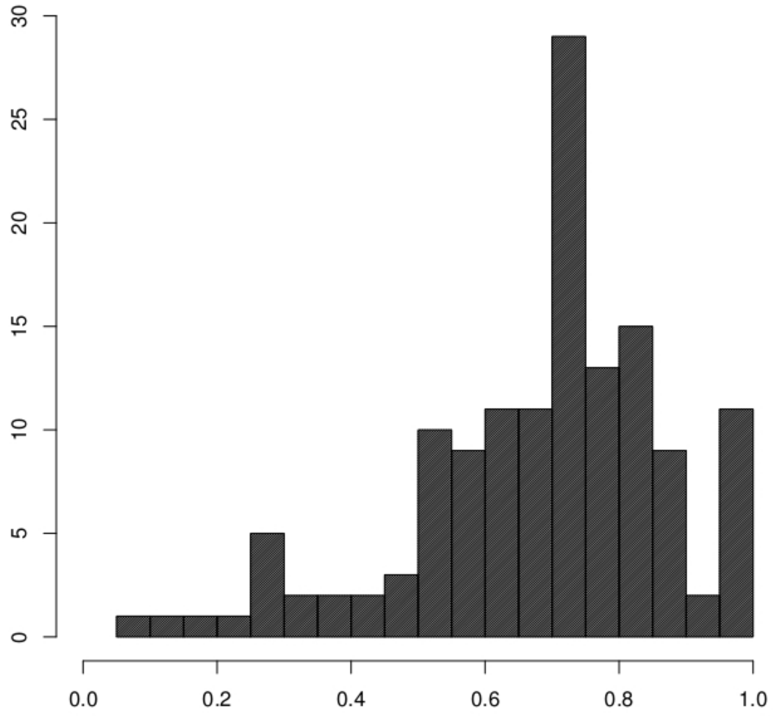


Figure 4.1: *Cell-type proportion coming from cell-type A*

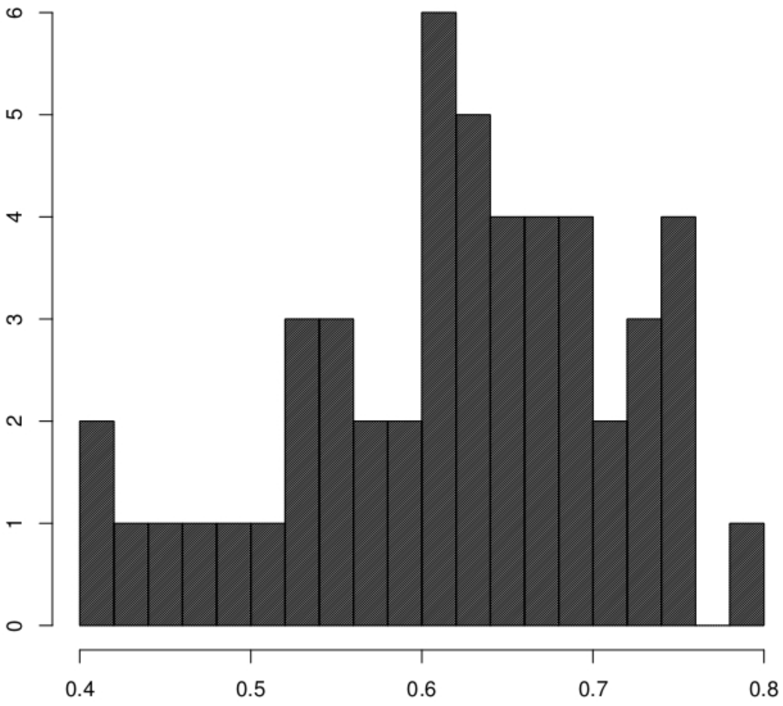


Figure 4.2: *Correlation coefficient between γ and γ_{init}*

Ind	Freq	Median	StDev	Ind	Freq	Median	StDev	Ind	Freq	Median	StDev
1	22	0.50232	0.06390	47	22	0.76970	0.03873	93	24	0.74723	0.03211
2	28	0.89223	0.03154	48	28	0.62814	0.03642	94	26	0.70954	0.02575
3	30	0.30261	0.06199	49	24	1.00000	0.01483	95	24	0.73487	0.03801
4	20	0.82665	0.03618	50	26	0.84807	0.03524	96	24	0.54838	0.03404
5	18	0.75715	0.03112	51	26	0.52494	0.04439	97	26	0.73453	0.04872
6	32	0.77365	0.04050	52	24	1.00000	0.00488	98	25	0.70569	0.04341
7	26	0.18666	0.06959	53	24	0.95672	0.04073	99	25	0.48093	0.06025
8	24	0.66470	0.04402	54	26	0.65449	0.04423	100	25	0.71126	0.05190
9	25	0.84728	0.03396	55	21	0.70131	0.03974	101	25	0.61982	0.03381
10	25	1.00000	0.01178	56	29	0.95856	0.02896	102	26	0.92673	0.03380
11	21	0.33294	0.05851	57	29	0.84793	0.02743	103	24	0.87815	0.04744
12	29	0.81797	0.03540	58	21	0.51317	0.05346	104	23	0.57093	0.03340
13	27	0.77784	0.03568	59	25	0.69708	0.04332	105	27	0.27046	0.07197
14	23	0.27446	0.07251	60	25	0.60430	0.03637	106	22	0.74910	0.03843
15	22	0.72933	0.03484	61	23	0.72979	0.03768	107	28	0.13606	0.04754
16	28	0.78616	0.03707	62	27	0.98667	0.02452	108	26	0.36042	0.05320
17	31	0.07329	0.05391	63	25	0.70967	0.03844	109	24	0.41036	0.06263
18	19	0.57529	0.06533	64	25	1.00000	0.00000	110	27	0.70320	0.03260
19	24	0.81418	0.03412	65	27	0.75711	0.03439	111	23	0.26841	0.07793
20	26	0.70919	0.04862	66	23	0.55492	0.05298	112	21	0.46997	0.04853
21	16	0.82229	0.04438	67	29	0.70133	0.03958	113	29	0.85441	0.03454
22	20	0.74431	0.03211	68	21	0.58986	0.03876	114	16	0.82128	0.03257
23	14	0.54205	0.06022	69	50	0.27609	0.07707	115	18	0.65121	0.02931
24	27	0.71412	0.04179	70	20	0.60101	0.04557	116	16	0.59291	0.03642
25	23	0.71132	0.04361	71	30	0.56308	0.03990	117	50	0.62378	0.03883
26	11	0.76323	0.02555	72	26	0.65164	0.02580	118	29	0.53579	0.05894
27	18	0.49045	0.03432	73	24	0.82454	0.04238	119	21	0.83020	0.04377
28	21	0.65206	0.06425	74	25	0.69363	0.04017	120	29	0.86144	0.04264
29	24	0.62245	0.03872	75	25	0.53874	0.05457	121	21	0.56987	0.05530
30	26	0.74936	0.03412	76	21	0.74674	0.02465	122	23	0.78717	0.04269
31	26	0.73117	0.03522	77	29	0.55800	0.04438	123	27	1.00000	0.00000
32	24	0.60008	0.04766	78	26	0.73672	0.04981	124	24	0.68857	0.03457
33	24	0.85578	0.03189	79	24	0.50899	0.05095	125	26	0.71067	0.04548
34	26	0.76069	0.03312	80	25	0.96298	0.03070	126	25	0.83901	0.03703
35	22	0.72936	0.03365	81	25	0.84641	0.02929	127	25	0.59067	0.03601
36	28	0.64256	0.04482	82	25	0.94350	0.02786	128	30	0.80786	0.03949
37	28	0.89287	0.03193	83	25	0.63727	0.03537	129	20	0.74533	0.02589
38	22	0.65929	0.03951	84	18	0.80737	0.03259	130	50	0.75709	0.03598
39	22	0.76279	0.03945	85	32	0.37257	0.06468	131	24	0.88766	0.04221
40	28	0.71761	0.03694	86	24	0.63281	0.03956	132	26	0.77906	0.03321
41	24	0.71366	0.03997	87	26	0.72892	0.04257	133	23	0.65932	0.04967
42	26	0.89553	0.03845	88	23	0.62599	0.03164	134	27	0.72244	0.03202
43	22	0.53481	0.03819	89	27	0.69953	0.04583	135	24	0.25481	0.06032
44	28	0.21875	0.03888	90	26	0.76213	0.03487	136	26	0.98296	0.03870
45	20	0.95064	0.03154	91	24	0.80577	0.03939	137	26	0.85445	0.02932
46	30	0.74963	0.03531	92	26	0.41258	0.06002	138	24	0.52435	0.04538

Table 4.5: *Individual mouse cell-type proportion estimates.*

Chapter 5

Discussion

5.1 Conclusion

The study of genetic variation is one of the exciting new frontiers in the study of biological organisms. The ability to understand how changes in the genetic code affect individual phenotypes is an essential part in understanding the underlying mechanisms which shape the living world. Mapping expression quantitative trait loci is one of the important ways in which we study genetic variation. As more and more eQTL studies are done, it becomes increasingly important to be able to cheaply and accurately estimate cell-type expression profiles. But the presence of cell-type heterogeneity reduces the ability to accurately estimate these profiles. This thesis presents a new methodology for estimating the cell-type specific expression profiles for a two cell-type model. It also estimates the proportion of the observed expression profile coming from each cell-type. The method employs an iterative least-squares regression approach for parameter estimation known as alternating-regression.

The first step of this process requires creation of initial cell-type proportion estimates for individuals of a dataset. Once we have these estimates we treat them as constants and estimate the gene-level means and slope paramaters of our two cell-types. Next, we treat these gene-level parameters as constants, and update our estimates of the cell-type proportions. Then, we fix the cell-type proportions and update the gene-level parameters, continuing back-and-forth indefinitely, until the parameter estimates pass some convergence criterion.

In order to test its accuracy, the parameter estimation algorithm was tested on simulated data. It was found that, for simulated data, we were able to use information contained in the first principal components loadings to construct initial cell-type parameter estimates which had 99% correlation with the actual values.

Under ideal simulation conditions (small random error introduced into the data), the parameter estimation algorithm was able to estimate all parameters with greater than 98% R^2 value. When more random error was introduced into the data, the median R^2 values for $\hat{\beta}$ and $\hat{\eta}$ dropped to 0.919 and 0.803, respectively (greater than 0.972 for the other parameters).

It was typical to be able to estimate parameter values for the primary cell-type more accurately than the secondary cell-type. This can be attributed to the distribution of the cell-type proportion parameter γ, which had most of its density favoring the primary cell-type.

Scatterplots of estimated parameters vs. the actual simulated values for β and η indicated the tricotomous distrubtion of the simulated data. Large absolute values of β, η were well estimated, and there were few true values of β, η close to zero incorrectly estimated as nonzero. The biggest problem that the estimation methodology had was with smaller (yet nonzero) absolute values of β, η being incorrectly estimated as zero (type I errors).

A real data analysis was performed on mouse hippocampus data using the proposed parameter estimation methodology. The dataset used was obtained from the GeneNetwork website, and was comprised of 138 BXD recombinant inbred strains of mice comprised of 69 distinct BXD strains. Since there were more mice than genotypes, this enabled us to estimate the parameters of many distinct permutations of mice.

Cell-type expression profiles for a two cell model were estimated using 50 different permutations of BXD mice. Parameter estimates for the full model correlated moderately with parameter estimates of μ and β for the simple regression model (median R^2 of 0.994 and 0.726, respectively). A histogram of estimated cell-type proportions for the individual mice had a density with highest concentration between values of 0.5 and 0.9.

The research in this thesis has shown that the alternating-regression methodology for parameter estimation of the two cell-type model can produce accurate estimates of cell-type specific expression profiles. These estimates greatly outperform estimates obtained from the standard simple linear regression model. For the researcher wishing to estimate cell-type specific expression profiles in a two cell-type setting, we present this methodology as a superior alternative.

Bibliography

[1] Avila VL (1995) “Biology: Investigating Life on Earth.” Massachusetts: Jones and Bartlett Publishers.

[2] Bioconductor: Open Source Software for Bioinformatics, cited 2012. Affymetrix Mouse Genome 430 2.0 Array annotation data. [Available online at `www.bioconductor.org/packages/2.2/data/annotation/html/mouse4302.db.html`]

[3] Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT (2005) “Mapping determinants of human gene expression by regional and genome-wide association.” Nature 437(7063), 1365-1369.

[4] Clarke J, Seo P, Clarke B (2010) “Statistical expression deconvolution from mixed tissue samples.” Bioinformatics 26, 1043-1049.

[5] Damerval C, Maurice A, Josse JM, de Vienne D (1994) “Quantitative Trait Loci Underlying Gene Product Variation: A Novel Perspective for Analyzing Regulation of Genome Expression.” Genetics 137, 289-301.

[6] Dubey RC (2006) “A Textbook of Biotechnology.” New Delhi: S. Chand

[7] Dietrich W, Katz H, Lincoln SE, Shin HS, Friedman J, Dracopoli NC, Lander ES (1992) “A genetic map of the mouse suitable for typing intraspecific crosses.” Genetics, 131:423-447.

[8] Erkkilä T, Lehmusvaara S, Ruusuvuori P, Visakorpi T, Shmulevich I, Lähdesmäki H (2010) “Probabilistic analysis of gene expression measurements from heterogeneous tissues.” Bioinformatics 26, 2571-2577.

[9] Ghosh D (2004) “Mixture models for assessing differentital expression in complex tissues using microarray data.” Bioinformatics 20, 1663-1669.

[10] Gosink MM, Petrie HT, Tsinoremas NF (2007) “Electronically subtracting expression patterns from a mixed cell population.” Bioinformatics 23, 3328-3334.

[11] Haley CS, Knott SA (1992) “A simple regression method for mapping quantitative trait loci in line crosses using flanking markers.” Hereditary 69, 315-324.

[12] Lander ES, Botstein D (1989) “Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps.” Genetics 121, 185199.

[13] Love JM, Knight AM, McAleer MA, Todd JA (1990) “Towards construction of a high resolution map of the mouse genome using PCR-analysed microsatellites.” Nucleic Acids Res 1990, 18:4123-4130.

[14] Leek JT, Storey JD (2007) “Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis.” PLoS Genet, 3(9): e161.

[15] Lu P, Nakorchevskiy A, Marcotte EM (2003) “Expression deconvolution: A reinterpretation of DNA microarray data reveals dynamic changes in cell populations.” Proceedings of the National Academy of Sciences of the United States of America 100, 10370-10375.

[16] Pierce JL, Lu L, Gu J, Silver LM, Williams RW (2004) “A new set of BXD recombinant inbred lines from advanced intercross populations in mice.” BMC Genetics 5:7.

[17] Pickrell JK Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK (2010) “Understanding Mechanisms Underlying Human Gene Expression Variation with RNA Sequencing.” Nature 464, 768-772.

[18] Plomin R, McClearn GE, Gora-Maslak G, Neiderhiser JM (1991) “Use of recombinant inbred strains to detect quantitative trait loci associated with behavior.” Behavior Genetics 21-2, 99-116.

[19] Shena M, Shalon D, Davis RW, Brown PO (1995) “Quantitative monitoring of gene expression patterns with a complementary DNA microarray.” Science 270, 467470.

[20] Stuart RO, Wachsman W, Berry CC, Wang-Rodriguez J, Wasserman L, Klacansky I, Masys D, Arden K, Goodison S, McClelland M, Wang Y, Sawyers A, Kalcheva I, Tarin D, Mercola D (2004) “*In silico* dissection of cell-type-associated patterns of gene expression in prostate cancer.” Proceedings of the National Academy of Sciences of the United States of America 101, 615-620.

[21] The National Institute of Health, cited 2012. The Genotype-Tissue Expression Project. [Available online at `http://www.genome.gov/gtex/`]

[22] The University of Tennessee, cited 2012. GeneNetwork: *bxd.geno*. [Available online at `http://www.genenetwork.org/dbdoc/BXDGeno.html`]

[23] The University of Tennessee, cited 2012. GeneNetwork: *GN110*. [Available online at `http://www.genenetwork.org/dbdoc/HC_M2_0606_R.html`]

[24] The National Institue of Health, cited 2012. Genetics Home Reference. [Available online at `http://ghr.nlm.nih.gov/handbook`]

[25] The University of Ghent, cited 2012. Homepage of Andy Vierstraete [Available online at `http://users.ugent.be/~avierstr/`]

[26] Venet D, Picasse F, Maenhaut C, Bersini H (2001) “Separation of samples into their constituents using gene expression data.” Bioinformatics 17, S279-S287.

[27] Veyrieras J-B, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK (2008) “High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation.” PLoS Genet 4(10): e1000214.

[28] Wang M, Master SR, Chodosh LA (2006) “Computational expression deconvolution in a complex mammalian organ.” BMC Bioinformatics 7, 328.