To the Graduate Council:

I am submitting herewith a thesis written by Adam Tebbe entitled "Low-Dose Radiation Effects on Variation in Gene Expression, With Potential Application to Network Construction." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Life Sciences.

Brynn Voy
Major Professor

We have read this thesis
and recommend its acceptance:

Arnold Saxton

Russell L. Zaretzki

Accepted for the Council:

Anne Mayhew
Vice Chancellor and
Dean of Graduate Studies

(Original signatures are on file with official student records.)

Low-Dose Radiation Effects on Variation in Gene Expression, With Potential Application to Network Construction

A Thesis Presented for the Master of Science Degree
The University of Tennessee, Knoxville

Adam Tebbe
December 2005

Abstract

The molecular responses to radiation exposure have been an avidly pursued research topic for many years now. Most of this effort has been focused on large doses of radiation, such as those experienced during a nuclear explosion or cancer treatment. Of equal importance though, are the effects of low doses of radiation, which have received much less attention. This study attempts to analyze the effects of low-dose radiation exposure on different inbred mice strains, each of which represents a unique genetic constitution. The genetic background was found to have a dramatic effect on the variability of gene expression of the irradiated mice, with anywhere from 950 to 6900 genes significantly variable. This effect was found to be more extreme in some strains compared to others, as well as being much more prominent in skin tissue than spleen tissue. Additionally, 200 to 3500 of the genes found to have variable expression were only variable in one of the inbred strains, which again show that the genetic background has an impact on different responses to radiation exposure. It was also found that performing microarray hybridizations using RNA that comes from multiple animals dramatically increases the variability of gene expression, when compared to RNA that comes from a single animal.

Pathway analysis and network construction methods have become a popular topic of interest in the bioinformatics community, which has been a result of the large volumes of data being produced through high-throughput genomics experiments. One popular method for building such pathways are through the use of Bayesian Networks. Many questions remain unresolved in their application though, such as how many experiments are required before an accurate model can be produced, and how accurate is that model. Based upon the simulations run, a minimum of 300 data points is required to obtain a graph that resembles the true structure, while more than 300 data points continues to improve the accuracy. The same number of data points also consistently produced the highest scoring graph, and allowed for the identification of modules within the graphs. The accuracy of this decreased though as the size of the graphs increased. Therefore, in consideration of applying such a method to the biological domain, a minimum of 300 experiments should be included before attempting to build a small network.

Table of Contents

Chapter 1: Introduction & background…….....1
I. Introduction 1
II. Low-Dose Radiation 2
III. Microarray Normalization 5
IV. The Analysis of the Variability of Gene Expression 6
V. Clustering 9
VI. Bayesian Networks 12

Chapter 2: Variability of gene expression in inbred mice strains in…….....17
response to low-dose radiation exposure
I. Abstract 17
II. Introduction 17
III. Methods 18
IV. Results 27
V. Discussion 92

Chapter 3: Bayesian network analysis……...…97
I. Abstract 97
II. Introduction 97
III. Methods 98
IV. Results 105
V. Discussion 125

Works Cited…….127

Vita……...135

List of Tables

Table 2.1: Number and type of arrays used. 19
Table 2.2: Number and type of arrays used for the study of a single animal versus a pooled sample. 20
Table 2.3: The number of genes with a significant variance ratio that are overlapping between strains. 35
Table 2.4: The number of genes with a significant variance ratio in the skin and spleen that are overlapping between strains. 36
Table 2.5: The number of genes differentially expressed between strains. 40
Table 2.6: The number of genes differentially expressed between the control and treatment groups. 41
Table 2.7: Summary of results from the analysis of different sets of genes. 84
Table 3. 1: Definition of metrics used for assessing accuracy of the topological structure of a learned graph. 101
Table 3.2: Metrics of accuracy for a graph with 5 nodes. 106
Table 3.3: Metrics of accuracy for a graph with 8 nodes. 107
Table 3.4: Metrics of accuracy for a graph with 10 nodes. 108
Table 3.5: Metrics of accuracy for a graph with 25 nodes. 109
Table 3.6: Average features matrix of graph with 5 nodes from 5000 data points. 114
Table 3.7: Results of features simulation of graphs with 5 nodes. 115
Table 3.8: Average features matrix of graph with 8 nodes from 5000 data points. 116
Table 3.9: Results of features simulation of graphs with 8 nodes. 117
Table 3.10: Average features matrix of graph with 10 nodes from 5000 data points. 118
Table 3.11: Results of features simulation of graphs with 10 nodes. 119
Table 3.12: Results of features simulation of graphs with 25 nodes. 121
Table 3.13: Average features matrix of graph with 25 nodes from 5000 data points. 122

List of Figures

Figure 1.1: Mouse family tree. 4
Figure 1.2: A directed acyclic graph, with the conditional independence relations in P(W,X,Y,Z). 13
Figure 2.1: Example scatter plot of pre- and post-within-print-tip group normalization. 22
Figure 2.2: Example boxplots of pre- and post-multiple slide normalization. 23
Figure 2.3: Number of genes in each strain and tissue with a statistically significant variance ratio ($p < .05$). 28
Figure 2.4: REML estimate between strains vs. within strain for the control groups. 29
Figure 2.5: REML estimate between strains vs. within strain for the treatment groups. 30
Figure 2.6: The number of significantly differentially expressed genes between the control and treatment groups in each strain and tissue. 31
Figure 2.7: The number of significant differentially expressed genes between strains for each tissue in the controls group. 32
Figure 2.8: The number of significant differentially expressed genes between strains for each tissue in the treatment group. 33
Figure 2.9: The number of genes with a significant variance ratio that were unique to a particular strain within each group. 37
Figure 2.10: The number of genes significantly differentially expressed between the control and treatment groups, which were unique to a particular strain in each of the two tissues. 38
Figure 2.11: The number of genes significantly differentially expressed between the control and treatment groups and also had a significant variance ratio, that were unique to a particular strain in each of the two tissues. 39
Figure 2.12: Hierarchical clustering of arrays. 42
Figure 2.13: Hierarchical clustering of strains with arrays averaged. 43
Figure 2.14: Ordered distance matrix and bootstrap reappearance proportions of genes with a significant variance ratio that were common between 5 of 6 strains in the skin treatment group. 45
Figure 2.15: Gene density plots of genes with a significant variance ratio and were common between 5 of 6 strains in the skin treatment group. 46
Figure 2.16: Ordered distance matrix and bootstrap reappearance proportions of genes with a significant variance ratio in 5 of 6 strains in the spleen treatment group. 48
Figure 2.17: Gene density plots of genes with a significant variance ratio in 5 of 6 strains in the spleen treatment group. 49
Figure 2.18: Ordered distance matrix and bootstrap reappearance proportions of genes with significant variance between the strains, but not within the strains in the skin treatment group. 50
Figure 2.19: Gene density plots of genes with significant variance 51

between the strains, but not within in the skin treatment group (within strain variance < 20,000 and between strain variance > 100,000).
Figure 2.20: Gene density plots of genes with significant variance between the strains, but not within in the spleen treatment group (within strain variance < 20,000 and between strain variance > 100,000). 53
Figure 2.21: Ordered distance matrix and bootstrap reappearance proportions of genes with a significant variance ratio in the treatment group, but not in the control group in the skin tissue, in 5 of 6 strains. 54
Figure 2.22: Gene density plots of genes with a significant variance ratio in the treatment group, but not in the control group in the skin tissue, in 5 of 6 strains. 55
Figure 2.23: Ordered distance matrix and bootstrap reappearance proportions of genes with a significant variance ratio in the treatment group, but not in the control group in the spleen tissue, and common between 5 of 6 strains. 56
Figure 2.24: Gene density plots of genes with a significant variance ratio in the treatment group, but not in the control group in the spleen tissue, in 5 of 6 strains. 57
Figure 2.25: Gene density plots of genes with a significant variance ratio in the control and treatment groups, and was differentially expressed between these groups in the skin tissue. 59
Figure 2.26: Ordered distance matrix and bootstrap reappearance proportions of genes with a significant variance ratio in the control and treatment groups, and was differentially expressed between these groups in the skin tissue. 61
Figure 2.27: Ordered distance matrix and bootstrap reappearance proportions of genes that were significantly differentially expressed between the control and treatment groups, in all of the strains in the skin tissue. 62
Figure 2.28: Gene density plots of genes significantly differentially expressed between the control and treatment groups, in all of the strains in the skin tissue. 63
Figure 2.29: Ordered distance matrix and bootstrap reappearance proportions of genes that were significantly differentially expressed between the control and treatment groups, in all of the strains in the spleen tissue. 65
Figure 2.30: Gene density plots of genes significantly differentially expressed between the control and treatment groups, in all 6 strains in the spleen tissue. 66
Figure 2.31: Ordered distance matrix and bootstrap reappearance proportions of genes that had a significant variance ratio in one and only one strain in the skin control group. 67
Figure 2.32: Gene density plots of genes that have a significant variance ratio in one and only one strain in the skin control group. 68
Figure 2.33: Gene density plots of genes that have a significant variance ratio in one and only one strain in the spleen control group. 69
Figure 2.34: Ordered distance matrix and bootstrap reappearance proportions of genes that had a significant variance ratio in one and only one strain in the 71

spleen control group.
Figure 2.35: Ordered distance matrix and bootstrap reappearance proportions of genes that had a significant variance ratio in one and only one strain in the skin treatment group. 72
Figure 2.36: Gene density plots of genes that have a significant variance ratio in one and only one strain in the skin treatment group. 73
Figure 2.37: Ordered distance matrix and bootstrap reappearance proportions of genes that had a significant variance ratio in one and only one strain in the spleen treatment group. 74
Figure 2.38: Gene density plots of genes that have a significant variance ratio in one and only one strain in the spleen treatment group. 75
Figure 2.39: Ordered distance matrix and bootstrap reappearance proportions of genes that were significantly differentially expressed between the control and treatment groups in one and only one strain in the skin group. 77
Figure 2.40: Gene density plots of genes that are significantly differentially expressed between the control and treatment groups in one and only one strain in the skin group. 78
Figure 2.41: Gene density plots of genes that are significantly differentially expressed between the control and treatment groups in one and only one strain in the spleen group. 79
Figure 2.42: Ordered distance matrix and bootstrap reappearance proportions of genes that were significantly differentially expressed between the control and treatment groups in one and only one strain in the spleen group. 81
Figure 2.43: Ordered distance matrix and bootstrap reappearance proportions of genes that were significantly differentially expressed between the control and treatment groups, and had a significant variance ratio in both the control and treatment groups in one and only one strain in the skin. 82
Figure 2.44: Gene density plots of genes significantly differentially expressed between the treatment and control group, and had a significant variance ratio in both the control and treatment groups, in one and only one strain in the skin. 83
Figure 2.45: The number of significantly differentially expressed genes between the single and pooled samples of RNA in each tissue and strain. 86
Figure 2.46: Number of genes in each strain with a statistically significant variance ratio ($p < .05$). 87
Figure 2.47: REML estimate between strains vs. within strain for the single-animal RNA. 88
Figure 2.48: REML estimate between strains vs. within strain for the pooled RNA. 89
Figure 2.49: Gene density plots of genes with a significant variance ratio within the strains, but not between in the skin treatment. 90
Figure 2.50: Gene density plots of genes with a significant variance ratio within the strains, but not between in the spleen treatment. 91
Figure 3.1: Sample plot of MCMC accept/reject ratio. 100
Figure 3.2: Sample plot of the node degree distribution of the original and learned graphs. 102

Figure 3.3: Example histogram of BIC scores from the features simulation. Scores were binned into groups of 3. 103
Figure 3. 4: Screenshot of ALIVE's interface. 104
Figure 3.5: Average True Positives, True Negatives, False Positives, False Negatives of highest scoring graphs with a given number of nodes. 110
Figure 3.6: Average Empty Rows, True Edges, Zero Edges and Significant Edges of highest scoring graphs with a given number of nodes. 111
Figure 3.7: Average False Positive Rate, False Negative Rate, True Positive Rate and True Negative Rate of highest scoring graphs with a given number of nodes. 112
Figure 3.8: Average Positive Predictive Value and False Discovery Rate of highest scoring graphs with a given number of nodes. 113
Figure 3.9: Graph from features matrix with 5 nodes from 5000 data points. 114
Figure 3.10: Graph from features matrix with 8 nodes from 5000 data points. 116
Figure 3.11: Graph from features matrix with 10 nodes from 5000 data points. 118
Figure 3.12: Average highest score of graphs in features simulation. 120
Figure 3.13: Graph from features matrix with 25 nodes from 5000 data points. 123
Figure 3.14: Scoring metrics of features graph with 5,000 data points. 124

Chapter 1

Introduction & background

I. Introduction

The study of the effects of ionizing radiation is an important topic in relation to human health. Everyone is exposed to a low level of background radiation. Natural sources of radiation come from a variety of locations, such as exposure to radon, cosmic radiation and ingestion [1]. In addition to this, many medical procedures such as CT scans, cancer treatments, consumer products and medical X-rays expose many people to an increased dose of radiation. There are also potential catastrophic events, such as atomic bombs, terrorist "dirty bombs" or nuclear accidents, which could produce large-scale radiation exposure to populations of entire cities. Given the wide variety of sources of radiation, both natural and man-made, it is important to understand the effects radiation exposure has on living cells and the human body. Much of the research done to this point has focused on the effects of large doses of radiation, leaving the effects of smaller doses of radiation fairly unknown. In addition, as for most diseases or environmental exposures, genetic background significantly influences the response to radiation, including exposures to low doses. Therefore, the study presented here focuses on the effects of 10 cGy of radiation, which is considered a low-dose by the Department of Energy. The goal of the study is 2-fold: 1.) to identify the genes that respond to low dose radiation in vivo as a means to identify pathways that are activated/repressed in the cell; and 2.) to determine the extent to which genetic background alters the radiation response. The latter goal is addressed by using inbred strains of laboratory mice to model the genetic variability seen in human populations.

In the field of bioinformatics, much attention has been given recently to the construction of gene regulatory networks, and methods for the high throughput elucidation of genetic pathways. In light of this, the ultimate goal in the study of the effects of radiation would be to know which pathways and regulatory networks are affected by the radiation exposure and exactly which genes are changing and in what ways they change. The methods being used to produce such networks are not entirely accurate or reliable, nor do they measure all the factors and events that one might wish to know in experimentally determining these regulatory networks. Thus, achieving a goal of completely determining such networks remains elusive. However, methods for constructing and modeling networks are steadily improving. There is a need for assess the current methods of elucidating networks, in part to provide feedback on the future needs for better experimental data and better network modeling methods. Therefore, a second goal of this study is to implement a common network construction algorithm and determine its effectiveness if applied to the gene expression data used in this study.

This thesis is organized as follows: the remainder of Chapter 1 contains background information on different topics relevant to Chapters 2 and 3. It is grouped together by topic, each of which appears in the order they are encountered in the subsequent chapters. Chapter 2 discusses the analysis of the low-dose data, and the

variability of gene expression both between and within the different inbred mouse strains. Chapter 3 is an investigation into pathway construction using Bayesian Networks, and discusses the potential for the elucidation of modules of interacting genes from high throughput data.

II. Low-Dose Radiation

Ionizing radiation is radiation that has enough energy to displace electrons from molecules [1]. Those displaced electrons can subsequently cause significant damage to live cells. Differing amounts of ionizing radiation have disparate biological effects. In general, radiation exposure is separated into two different groups: low- and high-doses. Low-dose radiation is generally defined as being anywhere in the range of zero to 100 mGy of radiation [1]. In the United States, the average annual amount of background radiation exposure is 3.0 mGy [1]. On average, one in one hundred people would be expected to develop cancer due to a lifetime of exposure to low-doses of radiation, which 40 of those 100 would be expected to develop cancer from leukemia or some other cause [1]. Despite the fact that the biological effect of radiation exposure is a widely studied topic, there is little information known about the effects of exposure to low doses of radiation [2]. In general, knowledge of the effects of low doses of radiation is extrapolated from knowledge of the effects of high doses of radiation, such as from Japanese atomic-bomb survivors [3]. While this may be acceptable under some circumstances, there have been shown to be effects of low-dose radiation exposure that differ from those exhibited resulting from high dose exposure, and vice versa [2].

Several models of the effects of low dose radiation exist. The most popular and controversial is the linear nonthreshold model, which proposes that low-dose radiation is just as harmful per gray as high-dose radiation [3]. Because of the complex biological effects of radiation, data exists that supports other models as well. One such model is a threshold model that proposes that low-dose radiation is harmless below a certain level. For the most part, this model is supported by the epidemiological data mainly of the life span of atomic bomb survivors, which seems to indicate in some cases a linearity between dose and risk at low doses [3]. Another popular model is the adaptive response model, which states certain amounts of low-dose radiation may even be beneficial [2-4]. Typically, the adaptive response is induced with 1-100 mGy of γ-rays, which is 100-10,000 times larger than natural background. A fourth model is the bystander effect model, which proposes that low-dose radiation may be even more damaging than that predicted by the linear nonthreshold model. It has been suggested that irradiated cells may signal their distress to each other, possibly through direct cell-to-cell interaction or by molecules secreted into the medium [3].

It has been found though, that low-dose radiation suppresses the molecular responses which prohibit the genotoxic effect of subsequent irradiation [5]. This is known as the radioadaptive response, and although the mechanism responsible for this is not yet known it has been suggested that suppression of p53-mediated apoptosis [6] and the responsiveness of protein kinase C are involved [7]. In humans the issue of radiation response is of particular importance, since many health risks have been associated with low doses of radiation. Such risks include neoplastic diseases, somatic mutations that

may contribute to other illnesses (including birth defects and other problems) and heritable mutations that may increase the risk of diseases in future generations [8, 9].

All amounts of irradiation are known to initiate a plethora of signal transduction cascades responsible for maintaining cellular homeostasis and promoting interactions with neighboring cells. This has been found to cause large-scale changes in gene expression levels, with specific implications for DNA repair, cell cycle arrest, growth control, and cell signaling [10]. The dose of radiation has been shown to have a significant effect on the number of genes with modulated expression patterns, as well as affecting genes with different functional classifications [11].

Because of the significant potential effects of radiation exposure noted above, and the current lack of understanding of the biological mechanisms affected by low-dose radiation exposure, this is an extremely important topic to pursue. While ideally the effects on humans are the focal point of any such study, for obvious reasons this cannot be done in a controlled study. Therefore, inbred mouse strains are an obvious alternative model organism. There are more than 450 different inbred mouse strains that now exist and are commonly used in laboratory experimentation [12]. A family tree of inbred mouse strains appears in *Figure 1.1*, which represents 102 of the strains [13].
While the different inbred mouse strains represent genetically distinct populations, multiple animals from the same inbred strain can be considered to have identical genetic constitutions. It is widely known however, that there are many polymorphisms that exist both within and between the strains, which could come into play when considering the effects of low-dose radiation. It has been found that the mean number of polymorphisms between strains is 608 ± 136 SD [13]. It has already been shown that different mice strains have variable sensitivities to mutagens, such as ethyl nitrosurea (ENU). In particular, BALB/c and C3H strains are particularly sensitive to this mutagen [12].

Because of the known genetic variation between strains, it stands to reason that there would be variability in the levels of gene expression as well. This issue is of particular interest to this study, as this will directly relate to inter-individual differences in response to the low-dose radiation treatment being applied. While most authors focus only on calculating differential expression between a treatment and control group, it is of great interest to also assess the variability between genetically distinct individuals. Therefore, not only the effects a low-dose radiation treatment, but also a baseline measure of the variability of gene expression between inbred mice strains would be of great value to the scientific community studying this topic.

The ultimate goal in studying the effects of radiation exposure is to understand the affected pathways, and how the transcription of the involved genes is modified in response to the treatment. While many methods have been developed to predict pathways from gene expression data, this is still a subjective and imperfect process. Due to the volumes of gene expression data being produced worldwide, pathway prediction has become an increasingly popular topic in the bioinformatics community; however it is fraught with complications and difficulties, as well as each method having well-known drawbacks. Of particular interest to this study, is how the genes identified as being highly variable and/or differentially expressed between the different mouse strains are related in terms of affected pathways. To work towards this goal though, it is first necessary to study the effectiveness and accuracy of network construction methods.

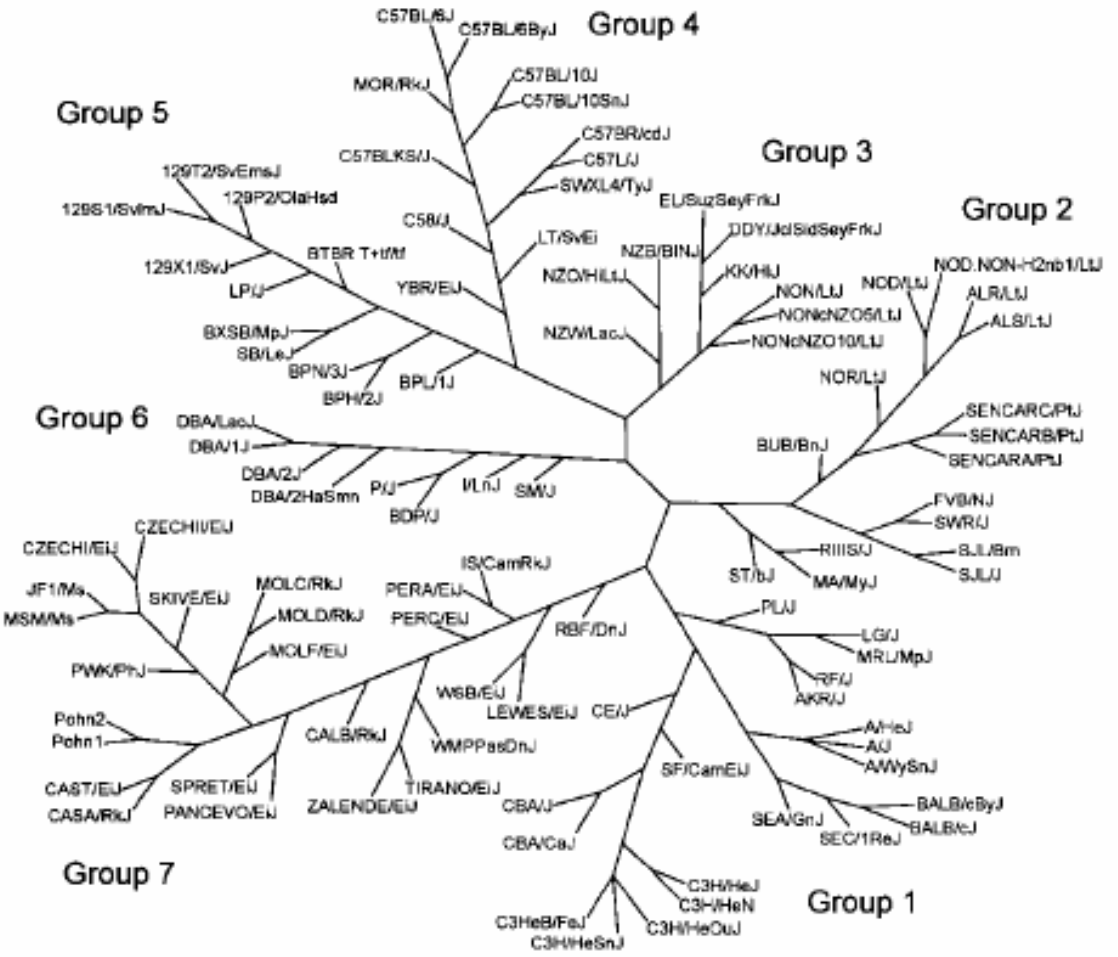


Figure 1.1: Mouse family tree. There are 7 groups of mouse strains. Group 1: Bagg albino; Group 2: Swiss mice; Group 3: Japanese and New Zealand inbred strains; Group 4: C57/58 strains; Group 5: Castle's mice; Group 6: C.C. Little's DBA and related strains; Group 7: Wild-derived strains. Branch lengths do not reflect evolutionary distances.

III. Microarray Normalization

The normalization of microarray data is an important step in the data analysis process. There are many sources of variation that exist in microarray experiments, which can have dramatic effects on measured gene expression levels [14]. Normalization is the process of removing technical sources of variation, so that the resulting data is comparable across different arrays in the same experiment, and biological differences can be more easily distinguished. There are many different ways that researchers have approached the normalization process. The types and amount of normalization required often depends upon the experimental design, and types of arrays used. Some of these methods are covered here, although this is not a comprehensive survey of all possible methods.

One of the earliest methods used for normalization was to have a set of "housekeeping" genes on an array, to which all genes were normalized. It was found though, that some of these genes were differentially expressed, so did not always exhibit the constant, basal level of expression that was expected [14, 15]. It was found to be very difficult to identify a set of genes whose expression does not change significantly under any conditions. A more accepted method was to use one array as a reference, and another which contained the treated samples [16]. In this way, all genes on the array had one measurement with the expected normal level of expression, and another measurement which was potentially different, as a result of some type of experimental treatment.

There are many factors that can contribute to systematic variability in two-color spotted microarray experiments. One such source of variation is dye bias, where the Cy3 and Cy5 dyes have differing intensities of fluorescence. Quite often, the intensity for the green dye is higher than that of the red dye, which can be caused by several different issues, such as the heat and light sensitivity or relative half-life of the dye, efficiency of incorporation, experimental variability in probe coupling and processing, or scanner settings [14, 15]. Many times, dye bias has appeared to be spot-dependent, which can be observed in plots of the log-ratio M vs. overall spot intensity A [14, 17]. To correct for this, a paired-slides normalization process can be used. This only applies to experiments where there is a dye-swap, in which there are two hybridizations for two mRNA samples, with dye-assignments being reversed in the second hybridization. One method to do this was proposed by Martin-Magniette et al. [18], where when the mean log-ratio was defined as

$$\log_2(R_1G_2/G_1R_2) \qquad \textit{Formula 1.1}$$

where R_1 is the intensity of the red dye on the first array, G_2 is the intensity of the green dye on the dye swap, G_1 is the intensity of the green dye on the first array, and R_2 is the intensity of the red dye on the dye swap. It was shown that using *Formula 1.1*, no genes were differentially expressed in a self-self hybridization [18]. Even using a dye-swap though, it is still possible that some variance exists, which cannot be removed. This can cause an increase in the false negative rate in an experiment, and subsequently decrease the power of detection [18, 19].

Another factor is a systematic variance of spots on the arrays as a result of differences among print-tips. Such differences could include differences in the length or opening of the tips, and deformation after extended periods of printing. Such variation is corrected through a within-print-tip group normalization method, which is defined in *Formula 1.2* below,

$$\log_2 R/G \rightarrow \log_2 R/G - c_i(A) = \log_2 R / (k_i(A)G) \qquad \textit{Formula 1.2}$$

where $c_i(A)$ is the LOESS fit to the M vs. A plot for the i^{th} grid on the array, and R and G are the red and green intensities, respectively [14, 15]. Any negative values for R or G will result in missing values for M and A, so are removed from further processing [20]. The LOESS scatter plot smoother performs a robust locally linear fit to the M-A plots, and is not affected by a small percentage of differentially expressed genes, which appear as outliers in the M-A plots [15, 21, 22]. A second source of variability within a slide is the ratios of expression for genes with low intensities, which can also vary substantially across experiments [23]. LOESS corrects for these effects as well though.

Another systematic source of variation lies in the differences inherent in using multiple arrays in an experiment. Following the within-slide normalization, all of the normalized log-ratios should be centered around zero, regardless of the method used to do this. Different slides can have different spreads in their log-ratios, which may need to be adjusted for through scaling [14]. A difference in the spread can result in the misidentification of genes that are differentially expressed in the treatment group compared to the control group [15]. Multiple-slide normalization attempts to correct for different sample variances in the log ratios across slides. These adjustments are necessary, so that the relative expression levels from one slide do not dominate the average relative expression levels across replicate slides [15].

A final source of systematic variation lies in the errors that result from the statistical tests used to identify differential expression. There is a constant tendency to over- and under-estimate true expression values, and cannot be removed through replicate analysis, since they are often highly reproducible [24]. This sometimes occurs with a dye-gene interaction effect in a model. One method that attempts to correct this is the use of a dye swap, although it must be noted that some errors are inherent in the statistical analysis used [24].

IV. The Analysis of the Variability of Gene Expression

Variation could be defined in terms of variation between replicate arrays of the same sample, or between replicate arrays of multiple samples from the same organism, or replicate arrays from multiple organisms. Of particular interest though, is the amount of variation that exists between animals in the same strain of a species, and the amount of variation that exists between different species of the same genus. Oleksiak et. al. studied this in different species of *Fundulus* in 2002, using two different geographically distinct populations of the fish [25]. They applied ANOVA methods to $\log_e$ normalized data, and found that 161 genes (18%) were significantly different between individuals within the same population ($p < 0.01$). For these genes, the difference in expression between two

individuals within the same population typically varied by a factor of 1.5, but varied by a factor of 2.0 or more for many genes [25]. From this, it can be concluded that two individuals within the same population can have a large difference in their levels of expression of the same gene. Oleksiak et. al. found that most of the variation between populations of fish is a result of the variation present within a population. This is in line with the neutral theory of evolution, which states that the variation between populations is a positive function of the variation within populations [25].

This study was followed up in 2005 with another study by the same group, where they investigated cardiac metabolism patterns of mRNA expression using microarrays [26]. They found that in *Fundulus heteroclitus*, which is a natural outbred population of fish, metabolism differed in individuals by greater than a factor of two. The expression levels of 94% of the genes measured were significantly different ($p < 0.01$), as determined by a nested ANOVA procedure with a Bonferroni correction [26]. It was also found that the statistically significant differences in gene expression were not due to one or a few individuals, or to experimental procedures. While there was a large amount of variation within populations, it was found that the expression of only 10% of genes was significantly different between populations, which provides additional evidence for the substantial variation in gene expression within a species [26].

A third study on the teleost fish was performed by Whitehead et. al [27]. This group found that 92 genes (48%) were differentially expressed among individuals within a population/tissue group and that inter-individual differences ranged over fivefold. They also found 146 genes (76%) that were differentially expressed among the brain, heart and liver, as determine by an ANOVA test ($p < 0.05$) with Bonferroni correction. Surprisingly though, only 31% of the tissue-specific differences was present in all three populations of fish tested [27]. Additionally, only 6 genes (3%) had differential expression among populations ($p < 0.05$), although this was cited to likely be an underestimate [27].

Studies comparing gene expression between strains have been performed in the two most popular model organisms for genetic dissection: *Drosophila* and *Saccharomyces*. It was discovered by Rifkin et. al. that differences in gene activity in *Drosophila* follow a phylogenetic pattern, and that 27% of all of the genes in the genome differ in their developmental expression between at least two strains or species [28]. For each gene in each strain, the differential expression along with confidence intervals was estimated by a general linear model and bootstrap randomization. Rifkin et. al. found 2,193 genes that were highly variable through this method. Jin et. al. performed analysis of *Drosophila* as well, through the use of sequential ANOVAs [29]. This group studied the effects that sex, genotype and age had on transcriptional variance. They determined that half of the genes exhibited sex-biased expression and 25% of genes may be genotype-biased. The average contributions of sex and genotype were determined to be 60% and 33%, respectively [29]. Studies in yeast by Brem et. al. suggested that nearly half of the yeast genome (2698 out of 6215 genes) is differentially expressed, as determined by the Wilcoxon-Mann-Whitney test with bootstrap correction [30].

Similar studies have been completed in the mouse as well, although very little is known about the normal amount of variation present [31]. Pritchard et. al found that 102 of 3,088 (3.3%) kidney genes, 62 of 3,252 (1.9%) testis genes, and 21 of 2,514 (0.8%)

genes in the liver had statistically significant ($p < 0.05$) variable expression, as determined by an ANOVA [31]. In this study, six normal male C57Bl6 mice were used, and the variance across these six mice was compared to the variance among four replicate experiments performed on each of the three tissues. Overall, they found that the kidney and testis had larger F values than the liver, which suggests a higher amount of gene expression variability in these tissues. Pritchard et. al. also considered whether or not highly expressed genes will tend to have higher F values as a result of a greater variance, but found that by plotting the F value vs. the intensity there was little correlation between these variables [31]. The majority of the statistically significant variable genes had an average intensity value near the mean of all expressed genes on the array.

While the study by Pritchard et. al. was informative about a basal amount of variation in an inbred mouse strain, it only considered one strain and did not address the amount of variation between different isogenic strains. It is well known that single nucleotide polymorphisms (SNPs) play an important role in differentiating between strains, and most likely contribute to the variability of gene expression as well [32, 33]. It has been found that up to 6% of the estimated 35,000 mouse genes contain blocks of SNPs, which contribute to functional regulatory variation [33]. Turk et. al. studied the variation of gene expression between several inbred mouse strains using an ANOVA method. Only two mice from each strain were tested, and significance levels between them were determined using a hierarchical t-test. False discovery rate was applied with a $p < 0.1$ level of significance, and found 88 out of 6,144 (1.4%) expressed genes were differentially expressed between strains [32]. A high correlation between gene expression levels for the two samples of each strain was found, which indicates a low amount of variation within the strains. The only tissue used in this study was the hindlimb muscle tissue. Overall, Turk et. al. estimated that in this tissue, approximately 1.4% of the expressed genes show a differential expression between the different strains. While the study by Turk et. al. only considered one tissue, a study by Pavlidis et. al. analyzed six different brain regions in two mouse strains [34]. A two-factor ANOVA and feature selection was performed to identify strain- and/or region-specific variation in gene expression. Pavlidis et. al. identified 65 genes that showed strain-specific variation ($p < 10^{-5}$) in expression across all brain regions tested, and 600 genes (6%) that showed region-specific variation ($p < 10^{-5}$) [34]. A similar study was completed on the same data set by Sandberg et. al., who found that approximately 1% of expressed genes are significantly different between strains in at least one region of the brain [35]. In this study, genes from one region were compared with all other brain regions using a Student's t-test ($p < 0.05$) to detect significant differences. A total of 73 genes were found to be differentially expressed in at least one brain region between the two strains (1.0%). No genes were found to be uniquely expressed in only one brain region, and only 2 genes were found to be differentially expressed in comparisons of all duplicate brain samples from the same strain [35].

The variation of gene expression has been addressed in humans as well by Cheung et. al., who assessed natural variation in lymphoblastoid cells [36]. They completed a genome-wide assessment of gene expression in humans by calculating a variance ratio, or F statistic for each gene by dividing the variance of the expression levels among individuals by the variance within microarray replicates. The variance

ratios of 813 genes ranged from 0.4 to 64, with a median value of 2.5 and the variance ratio within individuals ranged from 0.01 to 0.18. Cheung et. al. found that the 40 most variable genes were not clustered in any chromosomal locations, but seemed to be randomly scattered throughout the genome. They also assessed the functional classification of the 100 most variable and 100 least variable genes through the use of the Gene Ontology (GO). A chi-square analysis of the 8x2 contingency table was done to identify significantly different functions ($p < 0.025$) for variable versus non-variable genes. Many different categories were found to be significant in this comparison [36]. Even though there were no significant chromosomal clusters of highly variable genes, the authors concluded that some of this variation was genetic.

One interesting study was completed by Enard et. al. [37], who considered the intra- and inter-specific variation in primate gene expression patterns. All possible comparisons between six human, six chimpanzee and two orangutan animals in both the brain and liver were performed. Even though the test performed was more of a phylogenetic analysis, which was performed by calculating a pair-wise distance between two samples by summing the base-two logarithms of the absolute values of the fold change of each gene, there were still some relevant inferences drawn. Enard et. al. found that the amount of divergence of *Mus spretus*, *Mus caroli* from *Mus musculus* is in the same order of magnitude as that of chimpanzees (1.08%) and orangutans (2.98%), from humans [37].

An important factor to consider when investigating the variability of gene expression in different strains or species is the underlying genetic variation. Chesler et. al. estimated that with three replicate arrays per strain, that the amount of variance accounted for by strain has a median of 11% and is as high as 78% across all transcripts [38]. Differences in mRNA expression can be due to any number of factors, which not only include genetic variation, but also complex interactions of environmental factors and cell-cell interactions. The genetic variation is due to differences produced by *cis*-acting polymorphisms, which are most commonly located in the gene's promoter region, and by *trans*-acting variants, which could be located anywhere throughout the genome [38]. Inbred mouse strains are particularly well known for having polymorphisms. In alignment of the C57BL/6J to sequences from other strains, Wade et. al. found that there are long segments of either extremely high (~40 SNPs per 10 kb) or extremely low (~0.5 SNPs per 10 kb) polymorphism rates [39]. Only one-third of the genome falls in the category of long SNP segments. In looking at 50-kb segments of DNA, the observed distribution of SNP rates was found to be noticeably non-random. This is in contrast to the regions of low polymorphism rates, where the estimated rate of SNPs is one per 20,000 bp [39]. Comparison of common laboratory strains with B6 found that two-thirds of the genome has very low polymorphism rates, which indicates that there are genomic regions in the which the strains share a very recent common subspecies origin [39].

V. Clustering

Microarray technology has advanced to a state where it now can be considered one of the primary tools for genomics research. Due to its high-throughput nature, and ease of experimental design, the technology has exploded throughout the research

community. While the experimental part of this technology has become fairly standardized, there is still no clear-cut optimal method of analysis of the resultant data. Tools to mine microarray data are still constantly being developed, although no perfect solution is in sight. Every researcher seems to have their own favorite method of analysis, which creates a major problem for the comparison of results across different studies. While the bioinformatics community is attacking this problem with vigor, constantly publishing new ways to analyze this type of data, there does not seem to be any mainstream methods or standards being developed. Because of this, the ultimate goal of the elucidation of gene networks through the use of microarrays is not yet realistic.

Much of what is being done in terms of mining microarray data consists of clustering data, or grouping genes together based upon similar patterns of expression. This attempts to uncover meaningful relationships hidden in large, multivariate data sets. Many different mathematical techniques have been applied to this end, with varying amounts of success. While this problem is a popular topic for mathematicians and statisticians, it has not yet been able to effectively incorporate biological knowledge on the nature of the genes being clustered. There is no standard method to interpret gene clusters for biological relevance, other than manual interpretation by biologists, which can be time-consuming and difficult depending on the size and number of clusters produced [40]. One increasingly common method in biology for assessing gene function is through the use of the Gene Ontology (GO), which assigns "a dynamic, controlled vocabulary that can be applied to all eukaryotes, even as knowledge of gene and protein roles in cells is accumulating and changing" [41]. GO is represented as a directed acyclic graph structure, which easily lends to a tree structure when applying a distance function [40]. In this way, it can easily be integrated by clustering algorithms, to assess biological significance.

Many different methods and classes of clustering algorithms exist in the literature. Some have proven to be more popular than others, mostly based upon the ease of implementation, and the number of citations in the literature to which new results can be compared. Clustering techniques can be divided into hierarchical and non-hierarchical methods. Hierarchical methods produce a group of nested classes, which can be easily visualized in a tree (this resembles a phylogenetic classification) [42]. Hierarchical clustering, specifically unweighted pair-group mean average (UPGMA) using a correlation distance has become the most common method, and the standard to which all other methods are compared against [43]. There are many different distance metrics that can be applied to hierarchical clustering, each of which will yield distinct results.

Clustering algorithms can further be divided into divisive or agglomerative techniques [42]. A divisive method begins by placing all of the genes into a single cluster, which is gradually separated into smaller clusters. Agglomerative methods are the opposite, in that each gene is initially its own cluster, and these smaller clusters are gradually concatenated together to form larger clusters. These algorithms can be further partitioned into supervised or unsupervised methods [42]. Supervised methods include previously discovered biological information about the genes, and use this information as a guide in the clustering process. Unsupervised algorithms are the more commonly used method, as they cluster based only upon the normalized microarray data.

The most commonly used unsupervised algorithm, and one of the first statistical tests to be applied to the gene expression clustering problem is hierarchical clustering. This algorithm produces a hierarchy of clusters, rather than a set number of clusters fixed in advance. Initially, each observation forms its own cluster, and then at each subsequent level, the two 'nearest' clusters are combined to form one bigger cluster. There are many different metrics used with hierarchical clustering, to determine how close one cluster is from another. The most common metric is the Euclidean distance, which is a generalization of the Pythagorean Theorem. An average-linkage metric is used to calculate the distance between clusters. One method of calculating the averages is the unweighted pair-group method average (UPGMA). In this method, the average distance is calculated from the distance between each point in a cluster, and all other points in another cluster. The two clusters with the lowest average distance are joined into a new cluster [42].

Another commonly used method of partitioning expression data is through k-means clustering. This method requires a fixed number of clusters in advance. The algorithm assigns the observations into k clusters in order to minimize the total within-class sum of squares. Initially, all genes are randomly assigned to one of the k clusters. Then an average expression vector is computed, and used to compute the distances between clusters. Genes are then moved between clusters, to minimize intra-cluster distance, and maximize inter-cluster distance. Moved genes remain in their new location only if they are closer than to their previous cluster [42]. The number of clusters required for this method is often determined by principal component analysis, or human intuition.

A popular variation to k-means clustering is Kohonen's self-organizing maps [44]. Self-organizing maps are a neural network-based approach, which assigns genes to one of k partitions on the basis of the similarity of its expression vector to a reference vector. The method of creating these reference vectors is what distinguishes this method from standard k-means clustering. Reference vectors are created by training random vectors in an iterative process that moves genes to a reference vector that is closest to the gene [42]. While this method has become popular in the bioinformatics community, it still presents the problem that k-means clustering has, in that it requires a predetermined number of clusters, which is usually not a realistic expectation.

To alleviate this problem, Wang et. al. modified this algorithm and created double self-organizing maps [45]. In this modification, each gene has two vectors, instead of one – a weight vector and a position vector. Each time a gene is moved, both vectors are updated, and the closer the weights are between two genes, then the closer their position vectors will be. The number of clusters could then be determined by plotting the position vectors, and visually inspecting the plots. This method was further modified by Ressom et. al. in an algorithm named adaptive double self-organizing maps [46]. This method performs clustering and cluster visualization simultaneously, thereby requiring no *a priori* knowledge about the number of clusters. It is adaptive, because it updates its free parameters during training, and it allows convergence of its position vectors to a fairly consistent number of clusters, provided that its initial number of nodes is greater than the expected number of clusters. This simplifies the clustering process from the user's standpoint, and removes any discrepancy in plot interpretation from double self-organizing maps.

A commonly used method for partitioning data is through the use of graph theory. Graph theory is a branch of mathematics and computer science that studies the properties of collections of nodes and vertices. A large portion of computer science is devoted towards this discipline, and has recently been applied to the problem of clustering expression data. One such method uses minimum spanning trees [47]. A tree is a common representation of a graph, and is the representation produced by a dendrogram, such as is done in hierarchical clustering. The use of minimum spanning trees converts a multidimensional clustering problem to a tree partitioning problem. Each cluster is represented as a subtree, which does not overlap the representing tree of any other cluster. In this method as with many of the clustering methods, a gene can be represented in only one cluster, which is not a realistic expectation in a biological setting.

Another class of unsupervised algorithms that are gaining some attention are fuzzy algorithms. These algorithms determine the relative likelihood of each gene belonging to each of a number of clusters, thus providing information about a gene's multi-functionality. They also inherently account for noise in the data, because they extract trends rather than precise values. While this is more meaningful biologically, since these algorithms don't use exact values, they are not as powerful statistically. Two such algorithms are: fuzzy adaptive resonance theory (ART) [48] and fuzzy k-means [49]. Fuzzy ART uses a neural network with back-propagation. Fuzzy k-means is an extension of the k-means heuristic described above. Since it is fuzzy, it does not required a predetermined number of clusters, but does risk the potential of getting stuck in a local minimum of a poor value of the expression vector.

A majority of publications concerning the analysis of microarray data utilize experiments that are done over a period of time. Since it is not as meaningful to assess the expression changes of a single experiment, expression changes over time are often compared to some basal expression level of the different genes of interest, which indicates whether the gene exhibits a response to the treatment applied. The time-course data from a single species is then combined into large matrices, and used as the basis for clustering. While for many applications this is an adequate experimental design, it does not take into account evolutionary conservation of function or regulation, and excludes the use of sequence analysis tools. The integration of other biological data such as this, with unsupervised clustering algorithms has not yet been addressed.

VI. Bayesian Networks

Bayesian statistics is a fairly new branch of probability theory, which has exploded into many different areas of research. This field has found particular application in the study of genetics data [50]. Applications in genetics have included population genetics, sequence analysis, quantitative trait loci (QTL) analysis [51], pedigree and multilocus linkage analysis [52] and the construction of phylogenetic trees [53]. One particularly popular application of Bayesian statistics is the analysis of microarray data, which includes analysis of differential expression, cluster analysis, and regulatory networks [54]. Of particular interest in the present study, is the use of Bayesian methods to construct gene regulatory networks. Bayesian networks have received a good deal of attention in the last few years [55-58], not only with constructing

regulatory networks, but also for their ability to incorporate proteomic data [59, 60]. Bayesian networks have been applied to both spatial and time-course data [58], in the form of static and dynamic Bayesian networks, respectively. One important question in building a Bayesian network though, is what is the minimal amount of data required to attain an accurate model? Building a network that includes thousands of genes from several dozen experiments is very difficult, since such a small number of arrays does not suffice to distinguish between true correlations and false positives [57]. So an important questions remains, what is the optimal number of experiments needed?

A Bayesian Network is a graph-based model of joint multivariate probability distributions that captures properties of conditional independence between variables [55]. Such networks represent the dependence structure between multiple interacting quantities (i.e. genes), and attempt to uncover the transcriptional events that regulate the expression of genes. Bayesian networks are mathematically defined strictly in terms of probabilities and conditional independence statements; however, this can easily be extrapolated to the notion of direct causal influence [55]. A Bayesian network is simply a graphical way to represent a particular factorization of a joint distribution [61]. Due to the inferred causality present in the Bayesian Network structure, they naturally suggest control points in signaling pathways, which is of great interest biologically [62].

A Bayesian network structure is simple to interpret: each node in the graph is conditionally independent from its non-descendents given its parents [61]. Generally, two disjoint sets of nodes A and B are considered conditionally independent given C, if C d-separates A and B. In other words, if along every undirected path between a node in A and a node in B, there is a node D, such that: (1) D has converging arrows and neither D nor its descendents are in C, or (2) D does not have converging arrows and D is in C [61]. In this way, visual inspection of the graph of a Bayesian network is simple, and it is not necessary to go through scores of graphs and more mathematically rigorous methods of evaluation. An example of these principles is depicted to the right in *Figure 1.2*. Node W is conditionally independent from X given the set C = {Y, Z}, since Y ε C is along the only path between W and X, and Y does not have converging arrows. From this graph though, it cannot be inferred that W is conditionally independent from X, given Z [61]. From this example, it is also obvious that a directed, acyclic graph is required. Each factorization computed in a Bayesian Network implies a strict ordering of the variables, so this therefore forces the constraint of directed acyclicity upon the model. However, this constraint is removed through the implementation of dynamic Bayesian Networks, which allow for loops, such as regulatory feedback loops [63]. Dynamic methods are ideal for handling time series data, as well as modeling stochasticity, prior knowledge and hidden variables [64].

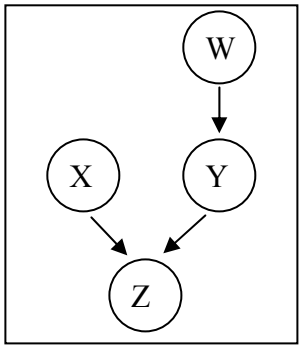


Figure 1.2: A directed acyclic graph, with the conditional independence relations in P(W,X,Y,Z).

There are two different types of variables that can exist in a Bayesian network, discrete and continuous. A

discrete variable can contain values from a finite set. If a variable is binary (2 possible values), then the table that specifies the probability of different possible values for the node will specify 2^k distributions. This immediately brings to light the complication in creating a Bayesian network, which is the amount of data required. In using a discrete variable, the number of free parameters is exponential in the number of parents [55]. The other type of variable is a continuous variable. Unlike discrete variables, when a node and its parents contain real values, there is no representation that can model all possible densities. Therefore, a natural choice to represent these multivariate continuous distributions would be to use Gaussian distributions. In a Bayesian Network, this consists of using linear Gaussian conditional densities [55]. In doing this, the node is distributed around a mean that depends linearly on the values of its parents. The variance of this normal distribution is independent of the parents' values. Both of these types of nodes can be utilized in a single Bayesian Network, which is called a hybrid network. This poses an important question of how to represent a conditional distribution for a continuous variable with discrete parents, and for a discrete variable with continuous parents [55]. To add to the complexity of variables in a Bayesian Network, variables can also be observed or hidden. A variable that describes an observed value is an information variable, while a variable describes an unobserved value is called a latent variable [65].

In learning a Bayesian Network, the most important question to consider is how to evaluate the accuracy of each possible network, and how to identify the optimal graph. One commonly used method is the Bayesian (or BDe) scoring metric. In this metric, the posterior probability of a graph given the data is defined as:

$$\begin{aligned} S(G : D) &= \log P(G \mid D) \\ &= \log P(G \mid D) + \log P(G) + C \\ P(D \mid G) &= \int P(D \mid G, \Theta)\, P(\Theta \mid g)\, d\Theta \end{aligned} \qquad \textit{Formula 1.3}$$

where D is the training set, G is a Bayesian Network structure, C is some constant independent of G and P (G) and P (Θ | G) are priors for G . This method includes a penalty for model complexity, thereby guarding against over-fitting the model [65]. The choice of priors determines the exact Bayesian score. Given a sufficiently large number of samples, those graph structures that exactly capture all dependencies in the distribution will receive, with high probability, a higher score than all other graphs. There are two problems with this though: in biological applications there often is not a sufficiently large number of samples, and the choice of the prior is of crucial importance, which is rarely known [55, 66].

An alternative metric was thus developed, which is a heuristic that does not require the existence of a defined prior. The Bayesian Information Criterion (BIC) score converges to a target model, however this does not ensure that the target model is the true data-generating distribution [67]. BIC is defined as:

$$BIC = -2 \ln(\zeta) + K \log(n) \qquad \textit{Formula 1.4}$$

where ζ is the maximum likelihood estimate, K is the number of estimable parameters in the approximating model and n is the sample size [67]. BIC is known to be biased when n is small, as an estimator of its target model.

Each prior in the network can be specified by two parameters: a prior network and an effective sample size [55]. The prior network reflects the belief on the joint distribution of the variables. The effective sample size reflects how strong the belief is in the prior network. Once the prior is specified and the data is given, learning consists of finding the graph that maximizes the score. This is known to be an NP-hard problem [55, 68], so heuristic methods are required. In most situations, graphs larger than 5 nodes are uncomputable, however Koivisto et. al. were able to compute graphs of 25 nodes using an algorithm based on dynamic programming [68]. Most commonly, local search procedures are implemented to do this, where one edge at a time is added, removed or reversed. Such methods include greedy hill-climbing, beam-search, stochastic hill-climbing, simulated annealing, K2, evolutionary algorithms and Markov Chain Monte Carlo [55].

Markov Chain Monte Carlo (MCMC) methods are an improvement over other Monte Carlo methods, such as importance sampling, where x random samples are drawn from the prior and the distribution on the hidden variables and then the samples are weighted by their likelihood, $P(y \mid x)$ where y is the evidence. Markov Chain Monte Carlo allows samples to be drawn from the posterior $P(X \mid y)$, even when the normalizing constant, $\int_x P(y \mid x) P(x)$ can't be computed [69]. MCMC is the most dominant method used for approximate inference of Bayesian graph structure. Typically, MCMC is applied to search over possible network structures, but it has recently been shown that it is sometimes more effective to search over all possible variable orders [70].

Bayesian Networks have recently become a popular tool for modeling genetic networks from noisy, high-throughput biological data. Bayesian Networks have the ability to simultaneously model non-linear combinatorial relationships, and robustly handle noisy data sets and guard against over-fitting [63, 71]. Since such networks capture linear and nonlinear interactions between groups of variables, and at the same time reduce the number of parameters required to describe the model, they are much more successful than other similar modeling approaches [62]. Bayesian Networks are also a great improvement over clustering methods, since Bayesian Networks can describe arbitrary control of gene expression, and are thus not limited to pair-wise interactions between genes [65]. In any experiment, only a subset of genes in a cluster may contribute to its expression signature, and different gene clusters may have similar signatures across the arrays, owing to either an overlap between the gene sets or co-regulation of non-overlapping gene sets [72]. Because of this, it is hard to conclude that such clusters actually correspond to pathways, as the expression profile is sometimes considered only a weak indicator that genes participate in the same pathway [73]. As such, many authors have considered incorporating alternative data sources, such as protein-protein interactions [60, 73, 74], genetic sequence data [75] and *cis*-regulatory elements [76], metabolic data [77], and human inference [64, 78, 79] to obtain more accurate network models. In the converse of this, Bayesian Networks have also been employed to predict protein-protein networks from genomic data [59].

Other methods for constructing Bayesian Networks from gene expression data have consisted of limiting the size of network being learned. In this way, the larger problem of network learning is broken down into smaller pieces, or modules, which can later be combined into larger graphical structures. One such method attempted to only learn the skeleton of the graph, which significantly restricted the search space of possible graphical structures [80]. Another such method determined that motifs were more reliable structures found in the expression data than entire networks, as determined by the scoring method used [81]. The identification of motifs, or features, in a larger graph has also been done with nonparametric bootstrapping [55] and greedy hill-climbing [82]. Clustering has also been attempted before network construction, to again attempt to limit the search space [83].

Chapter 2

Variability of gene expression in inbred mice strains in response to low-dose radiation exposure

I. Abstract

The molecular responses to radiation exposure have been an active research topic for many years. Most of this effort has been focused on large doses of radiation, such as those experienced during a nuclear explosion or cancer treatment. Of equal importance though, are the effects of low doses of radiation, which have received much less attention. This study attempts to analyze the effects of low-dose radiation exposure on different inbred mice strains, each of which represents a unique genetic constitution. The genetic background was found to have a dramatic effect on the variability of gene expression of the irradiated mice, with anywhere from 950 to 6900 genes significantly variable. This effect was found to be more extreme in some strains compared to others, as well as being much more prominent in skin tissue than spleen tissue. Additionally, 200 to 3500 of the genes found to have variable expression were only variable in one of the inbred strains, which again show that the genetic background has an impact on different responses to radiation exposure. It was also found that performing microarray hybridizations using RNA that comes from multiple animals dramatically increases the variability of gene expression, when compared to RNA that comes from a single animal.

II. Introduction

The response to radiation exposure is a topic that has been studied for many years, in relation to different radiation sources. Cells are known to exhibit varying responses to radiation exposure, depending upon the radiation dose, type of radiation, and genetic constitution. Since most research studying the effects of radiation exposure has focused on large doses, the cellular responses to low-doses of radiation are still fairly unknown. The goal of this study is to discover and characterize genetically inter-individual variation in response to low-dose radiation. In this study, inter-individual genetic variation is achieved through the use of inbred mouse strains as a model organism. It is believed that responses to low-dose radiation will differ among individuals because of differing genetic backgrounds, and because of this, there will be variation in the expression of genes that cooperate in the response pathways.

It has been known for more than 40 years that different strains of mice are differentially susceptible to the effects of radiation [84]. The BALB/c strain is well known to be sensitive to radiation, in particular it is known for being highly susceptible to radiation-induced mammary cancer and its genomic instability [84]. A major part of this sensitivity has been attributed to the deficiency in BALB/c to repair double-strand DNA breaks [84]. Low-dose radiation also causes differing immune responses in C57BL/6 and BALB/c, in that mitogen-induced proliferation of splenocytes in C57BL/6 is enhanced by

pretreatment with low-dose radiation, whereas it is suppressed in BALB/c [84]. Examples such as these attest to the variability of responses to low-dose radiation due to the genetic constitution.

Since responses to low-dose radiation treatment are not well known, it would not be a worthwhile endeavor to look at differences in the expression of specific genes (such as is done using an RT-PCR assay). Instead, high-throughput gene expression studies can be done to assay most of the mouse genome. This will allow for the characterization of many genes, both known and unknown. Of particular interest in this study, is not only how the expression differs in genes between the different strains, but also how much variability is present in the expression. It would be expected that since animals within the same inbred strain are almost genetically identical, that their response to radiation exposure would be identical as well. This can be assessed using a test similar to a one-way analysis of variance. Of related interest is how expression between the control and treatment groups differ, as well as how it differs between tissues.

III. Methods

Experimental Design:

Mice used in this study were all born and raised at the Oak Ridge National Laboratory William L. and Liane B. Russell Laboratory for Comparative and Functional Genomics. Six inbred mice strains were used for this study: A/J, C3H, B6.C, BALBxJ, Bl6Jx and DBA/2J. Each strain has a different genetic constitution from the others, while mice of the same strain can be considered to be genetically identical. In this way, it is possible to have replicate samples within the study, which come from the same genetic background. Treated mice were exposed to 10 cGy of X-rays, when they were 10 weeks of age. The mice were sacrificed 3.5 hours after exposure, and the tissue was stored at -80°C until used for RNA extraction. RNA was extracted from the spleen and skin tissues, converted to cDNA and hybridized to custom microarrays designed using the Mouse Oligonucleotide Library by Sigma-Genosys and Compugen. There were 26 slides that were printed using the 15,000 oligonucleotide (oligo) version of this mouse library, and the other 78 slides were printed with the 22,000 oligo version. The oligos were all 65 base pairs long. The 22,000 oligo version of this library included the entire 15,000 oligo version. Each of the arrays printed from the 15,000 oligo library contained 15,120 spots. There were three different types of spots on the array, with 144 spots that were left blank, 115 spots that contained only buffer, and 14,861 spots that contained oligos. Each of the arrays printed from the 22,000 oligo version of the library contained 23,232 spots. There were four types of spots on these arrays, with 865 spots left blank, 225 spots that contained only buffer, 96 spots that contained alien (unknown) oligos, and 21,997 spots that contained oligos.

A dye swap of each array was done as well, to account for incorporation and quantum characteristics of the two dyes. The number of arrays used in each tissue is listed in *Table 2.1* below. Each array was hybridized with a pool of two RNA samples, 1 of which was hybridized with Cy3, and the other with Cy5. For each array, there is a corresponding dye-swap, where the dye assignments are reversed. In *Table 2.1*, the dye

Table 2.1: Number and type of arrays

Strain	**Tissue**	**Number of Arrays**	**Array Type**
A/J	Skin	4	22K
		4	15K
	Spleen	8	22K
B6.C	Skin	8	22K
		6	15K
	Spleen	8	22K
BALBxJ	Skin	8	22K
		4	15K
	Spleen	8	22K
Bl6Jx	Skin	8	22K
		4	15K
	Spleen	4	22K
C3H	Skin	8	22K
		4	15K
	Spleen	6	22K
DBA/2J	Skin	4	22K
		4	15K
	Spleen	4	22K

Table 2.2: Number and type of arrays used for the study of a single animal versus a pooled sample.

Strain	Tissue	Number of arrays	Array type
B6.C	Skin	8	22K
	Spleen	8	22K
B6/J	Skin	8	22K
	Spleen	0	22K
C3H	Skin	6	22K
	Spleen	0	22K

swap is counted as a separate array, so there are essentially half as many arrays in the study as is listed in the table.
A second component of this study assessed the variability of gene expression of a RNA sample taken from a single animal, and compared it to a sample taken from a four-animal RNA pool. In this component, the pooled sample is labeled with Cy5, and the single-animal sample is labeled with Cy3. There is then a dye-swap done, such that the pooled-animal sample is labeled with Cy3 and the single-animal sample is labeled with Cy5. In *Table 2.2*, the dye swap also is counted as a separate array, so again there are essentially half as many arrays in the study as is listed in the table.

Spot Extraction:

ImaGene was used to extract the intensity measurements from the scanned images of the arrays, and to flag spots that may have been poor measurements. Empty spots were flagged if they had a low expression of were entirely missing. If the measure value was below a preset threshold, then a green "X" was marked on the image. Negative spots were flagged if the signal mean was lower than the background mean. With these spots, a green "-" was marked on the spot on the image. Poor spots were identified by comparing a background level to the local background level distribution over the entire image, as identified by a t-test. The resulting p-value was subtracted from 1, and labeled the "confidence in contamination presence". Spots with a low confidence were flagged. Lastly, some spots were manually flagged, which were not identified by the filters described.

Normalization:

Before any analysis, the raw intensity values were normalized. This was done through a series of steps in SAS, which attempted to remove within- and between-slide variability. The first step in the normalization process for this data set was to reset some of the flags in the skin data. During the spot extraction process, there were different technicians who used varying sensitivities for flagging genes in ImaGene. Because of

this, there were some arrays that had every single gene marked as being flagged, empty, poor or negative. Normally, any genes marked are simply removed from the data set, but since this would have removed entire arrays, a new procedure had to be implemented to reset these. Genes marked as being empty or poor were reset in the skin arrays, and then marked as empty if the signal mean was less than a cutoff value. The cutoff was set as the median intensity plus three times the standard deviation of the median of absolute deviation (madsd) of the buffer spots on the array.

After the flags were adjusted in the skin data, the signal of each spot was calculated as the signal mean minus the background mean. The signal was then transformed to a $\log_2$ scale. Following this, normality tests were applied to the data to identify outliers. An outlier was defined as having the absolute value of its signal minus the median value being greater than ten times the madsd.

Once the outliers were determined, these genes, along with the flagged genes were removed from the data set. There were initially 4,410,432 measurements in the data set. 7,597 measurements were outliers and removed, leaving 4,402,835 measurements. 1,347,131 spots were flagged and removed, which left 3,055,704 measurements in the data set. Following this, the normality of the coefficient of variation was assessed. While the removal of the flagged spots and outliers greatly decreased the skewness and kurtosis of the data set, some still remained. Despite this, it was assumed that this didn't have a large effect on the data, and the data was treated as being from a normal distribution (as recommended by Wolfinger et. al. [85]).

The second part of the normalization process was within-print-tip group normalization. Each grid on the array was printed with the same print-tips. There are some systematic effects that may exist on the array due to this, such as slight differences in the length, and there may be some deformation of the tips after excessive printing [14]. Additionally, printing the arrays one group at a time may create spatial effects on the slide. These effects are removed by using LOESS, which is a scatter plot smoother based on local regression. The log ratio, M (treatment minus control) is plotted versus the overall spot intensity, A (average of treatment plus control). The fraction of data used for smoothing at each point was 30%. Once this adjustment was calculated for each point in the scatter plot, half of the adjustment value was subtracted from the treatment point, and the other half was added to the control point. The M and A values were re-calculated, and 5,159 data points were removed due to missing M values. Scatter plots of the pre- (left) and post-normalized (right) data were created, an example of which is shown in *Figure 2.1*.

Following the within-print-tip group normalization, all of the log-ratios are centered around zero for each slide. With that being done, a control data set was created.

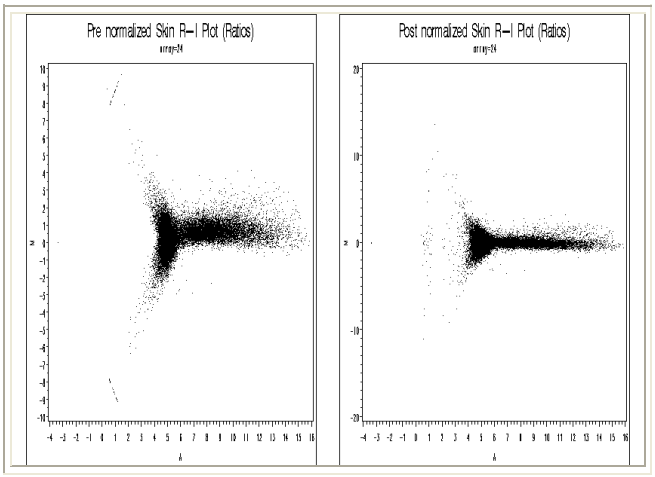


Figure 2.1: Example scatter plot of pre- and post-within-print-tip group normalization.

This was done by taking the untreated Cy3 array, and merging it with the untreated Cy5 array from the corresponding dye-swap. After the control data set was created, a multiple slide normalization method was applied, which allowed comparisons to be made between different experiments [14]. The data were normalized between slides using a mixed model, with the normalized log mean of the dye being the dependent variable in the model. Four of these models were used: one each for the skin and spleen controls, and one each for the skin and spleen radiation data. In each of these models, the only term in the model was the array. The model appears below in *Formula 2.1*,

$$d_{ij} = \varphi_i \qquad \textit{Formula 2.1}$$

where d_i is the log-normalized mean expression for dye j of gene i, and φ_I is the array. The result of this model was a log-normalized residual for each of the two dyes for each gene. Boxplots were created of the arrays, to show the effects of this part of the normalization. An example is shown in *Figure 2.2*.

The final part of the normalization process was paired-slide normalization (dye-swap). For the controls data set, there was no dye-swap, so the normalized residual values for the Cy3 and Cy5 measurements for each gene were simply averaged. The ratios data set was normalized according to the results of Martin-Magniette et. al [18], who showed through self-self hybridizations that using *Formula 2.2*, that no genes were differentially expressed,

$$r_i = R_1G_2/G_1R_2 \qquad \textit{Formula 2.2}$$

where R_1 is the Cy5 value from the treated part of the array, G_2 is the Cy3 value from the treated part of the array, G_1 is the Cy3 value from the untreated part of the array and R_2 is the Cy5 value from untreated part of the array.

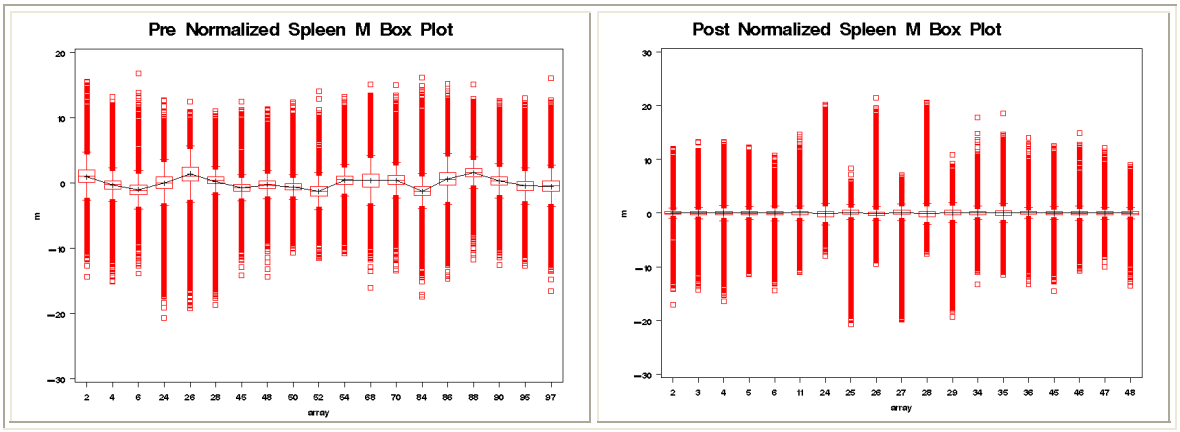


Figure 2.2: Example boxplots of pre- and post-multiple slide normalization.

The genes on the arrays were cataloged according to their GenBank ID name, so each gene with a GenBank ID had to be converted into an NCBI Gene ID, in order to map chromosomal locations and Gene Ontology terms. On the arrays, there were 18,591 genes that had NCBI Gene ID's, and 3,123 genes that did not.

Variance Ratio Calculation:

A variance component estimate was calculated for each gene in each strain using Proc Mixed in SAS. The estimation method used was to compute the restricted/residual maximum likelihood (REML) for each gene. This was done for each gene within each strain and for that gene between the different strains as well. The variance ratio is then defined as the REML estimate between the strains divided by the REML estimate within the strain. This method is essentially analogous to a one-way Analysis of Variance (ANOVA), which follows an F-distribution. Therefore, a p-value was calculated from an F-distribution. Genes with a p-value < .05 were considered significant.

Formula 2.3 was used to calculate the REML estimate within a strain for a gene,

$$r_i = \alpha_i + \beta_i \qquad \textit{Formula 2.3}$$

where r_i is the normalized residual for gene i, α_i is the strain and β_i is the gene id. In this model, the strain is a random effect, and strain and gene id are categorical independent variables in the model. *Formula 2.4* was then used to calculate the REML estimate between the strains for a gene,

$$\text{REML}_{\text{within}} = \beta_i \qquad \textit{Formula 2.4}$$

where $\text{REML}_{\text{within}}$ is the REML estimate for gene i within each strain, and β_i is the gene id. The variance ratio is then calculated as is shown below in *Formula 2.4* [86]:

$$\text{Var} = \frac{\text{REML}_{\text{between}}}{\text{REML}_{\text{within}}} \qquad \textit{Formula 2.5}$$

The p-values for the variance ratio of each gene in each strain was calculated using an F-distribution, where the number of degrees of freedom in the numerator was one minus the number of observations in *Formula 2.4*, and the number of degrees of freedom in the denominator was one minus the number of observations in *Formula 2.3*. Critical F-values were calculated with a .05 level of significance as well. Genes with a p-value < .05 were considered significant.

Differential Expression Between Treatment & Control Groups Calculation:

The differential expression of genes in the treatment group compared to the control group was calculated using Proc Mixed in SAS. A paired t-test was calculated using *Formula 2.6*, as follows:

$$r_i = \alpha_i + \gamma_i + \delta_i + \alpha_i * \gamma_i \qquad \textit{Formula 2.6}$$

where r_i is the normalized residual for gene i, α_i is the strain, γ_i is the treatment, δ_i is the array and $\alpha_i * \gamma_i$ is an interaction effect between the strain and treatment. In this model, the array is a random effect, and treatment, strain and array are categorical independent variables. This model was used for each tissue separately. A False discovery rate (FDR) correction was then applied to the p-values [87], to attempt to control the proportion of false-positives resulting from multiple comparisons. Genes with a p-value < .001 were considered significant.

Differential Expression Between Strains Calculation:

The differential expression of genes between strains was calculated using Proc Mixed in SAS. A paired t-test was calculated using *Formula 2.7*, as follows:

$$r_i = \alpha_i + \delta_i \qquad \textit{Formula 2.7}$$

where r_i is the normalized residual for gene i, α_i is the strain and δ_i is the array. In this model, the array is a random effect, and strain and array are categorical independent variables. This model was used for each tissue and treatment separately. A False discovery rate correction was not applied to the p-values, since such a small number of genes were significant (p-value < .05).

Clustering:

To cluster large sets of genes, the Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH) algorithm was implemented in the statistical package R. The HOPACH algorithm builds a tree of clusters, where the clusters in each level of the

tree are ordered based on the pairwise dissimilarities between cluster medoids [88]. HOPACH starts at the root node, and attempts to find the number of children for each node by alternating partitioning steps with collapsing steps. In this way, it is a combination of the two categories of clustering methods: agglomerative and partitioning, and does not require a predefined number of clusters.

HOPACH creates a hierarchical tree, which it then unfolds through a series of steps. First, the root node is partitioned, and the children in the next level are ordered deterministically using the same dissimilarity matrix that is used for clustering. Next, each of these nodes is partitioned, and its children are ordered. Before the next partitioning step, collapsing steps merge any similar clusters. The process is iterated until the main clusters are identified. The MSS criterion is used to automatically choose (i) the number of children at each node, (ii) which clusters to collapse, and (iii) the level containing the main clusters. Below the main clusters, the algorithm is run down without collapsing to produce a final ordered list. The algorithm is described in more detail in Pollard et. al [88].

HOPACH calculates clusters with both strict membership, and fuzzy membership. Fuzzy clusters allow for genes to have some amount of membership in multiple clusters, whereas clusters with strict membership permit genes to belong to one and only one cluster. Fuzzy clusters are determined by proportions of bootstrap reappearance. For all the applications of clustering in this study, clusters with fuzzy membership were used. Since HOPACH determines the number of clusters present in the data, there are many clusters of varying sizes created. Many of these clusters contain one to a few members, which for this application is too few. Therefore, clusters with less than 10 genes in them are combined together until a cluster of at least 20 is reached. Since the initial clusters are fuzzy, a gene is considered a member if it has any amount of membership greater than zero. If a cluster contains 10 or more genes, it remains as a separate cluster. Each cluster was then run through Gene Ontology analysis as described below.

There are two typical figures produced to show the effects of the HOPACH algorithm. The first is a plot of the ordered distance matrix and the second is a barplot of the bootstrap reappearance proportions for each gene and each cluster. In the plot of the ordered distance matrix, rows and columns of the dissimilarity matrix are ordered according to their final level in the HOPACH hierarchical tree. Red color indicates a small dissimilarity, and white color indicates a large dissimilarity, with red and white correspond to positive and negative correlation, respectively. Yellow color indicates no correlation. The second plot is a barplot of the bootstrap reappearance proportions for each gene and each cluster. Each cluster is represented by a different color, and each gene is represented by a narrow horizontal bar. The genes are ordered by HOPACH cluster, and then by bootstrap estimated membership within a cluster and plotted on the vertical axis. The length of each bar with a certain color is proportional to the percentage of bootstrap samples in which that gene appeared in the cluster represented by that color. If the bar is all or mostly one color, then the gene is estimated to belong strongly to that cluster. If the bar is many colors, the gene has fuzzy membership in all these clusters.

Chromosomal Clustering:

Clusters of genes on the mouse chromosomes were determined by plotting a density distribution of genes in the statistical package R. This method is distinct from previously used methods, which rely on a frequency count, or simple manual interpretation [89-92]. The chromosome number, and gene start site was first obtained for every gene in the set, and every gene in the mouse genome from GeneKeyDB [93]. A list was then compiled of the start site of every gene on every mouse chromosome. A density distribution was plotted for these, using the density function in R. This was also done for the start sites of every gene in the set. The distribution for each chromosome was plotted separately. On each of these plots in the Results section, the distribution for the mouse genome is plotted in blue, and the distribution for the gene set is plotted in green.

Then two permutations were done, to aid in determining regions of significance in the genome. The first permutation consisted of randomly sampling the same number of genes as was in the set for a given chromosome from all the genes on that chromosome in the mouse. This was done 1,000 times, and the density curve was calculated each time. The highest point in the density curve from each of the 1,000 samples was stored, and these points were then sorted. The 10^{th} highest point was then plotted in red on each density plot, which is a .01 level of significance. The second permutation consisted of randomly sampling the same number of genes as was in the set from the entire mouse genome. This was done 1,000 times as well, and a density curve was calculated for each chromosome. The highest point in the density curve for each chromosome was stored, and these were subsequently sorted. A .01 level of significance was used for these as well, which was obtained by multiplying the number of genes on a given chromosome in the sample times .01. This was then rounded to an integer value, and was used as the index to the vector of sorted points from the density curves. This was plotted in yellow on each chromosome. Any parts of the density curve that met or exceeded both lines from the permutations were considered significant.

Gene Ontology Analysis:

The analysis of a set of genes for enriched Gene Ontology (GO) terms was done using a modified Perl implementation of the GoTree Machine [94]. Gene Ontology categories with a significant number of genes were identified by comparing the distribution of genes in the interesting gene set in each GO category to those of all the genes on the array that had an NCBI Gene ID (the reference gene set). In calculating this, there are n genes in the set of interest, and N genes on the array with NCBI Gene ID's. For any given GO category X, a gene from the set is either present or not. If K out of N reference genes, and k out of n interesting genes are in category X, and if the n genes are effectively a random sample uniformly selected from N, then the expected value of k would be $k_e = (n/N)K$. If k exceeds the expected value, the category X is determined to be enriched, with a ratio of enrichment (R) given by $R = k/k_e$. Since genes can be selected only once, this is sampling without replacement and can be appropriately

modeled by the hypergeometric distribution. The significance of enrichment (P) for a given category is determined by *Formula 2.6* [94] below.

$$p = \sum_{i=k}^{n} \frac{\binom{N-K}{n-i}\binom{K}{i}}{\binom{N}{n}} \qquad \textit{Formula 2.6}$$

Genes Ontology categories with a p-value of 0.01 or less were determined to be significant. In addition, Bonferroni correction was used to adjust for multiple comparisons. Terms were only considered significant if they had three or more genes from the set of interest in them.

IV. Results

Figure 2.3 displays the number of significant genes in each strain and tissue for the control and treatment groups. There is a significant increase in the number of genes variable in the control versus treatment groups in each strain. In addition, the spleen tissue seemed to have much less variability in gene expression than did the skin tissue. *Figure 2.4* plots $REML_{between}$ vs. $REML_{within}$ for the controls, and *Figure 2.5* plots the same thing for the treatment groups. There were several extreme outliers, which were excluded from *Figure 2.5*. The lines in *Figures 2.4 and 2.5* are calculated linear regression lines. The $REML_{between}$ and $REML_{within}$ estimates are very small in the controls (*Figure 2.4*) when compared to the treatment groups (*Figure 2.5*). *Figure 2.4* also shows an interesting pattern, in that there seem to be a large number of genes with a large $REML_{between}$ and a small $REML_{within}$ estimate, and also many genes that seem to have a linear relationship between these two variables. In *Figure 2.5* there appear to be some genes that have a large $REML_{between}$ and a small $REML_{within}$ estimate, some genes where there is a linear relationship between these two variables, while many genes that have a large $REML_{within}$ and a small $REML_{between}$ estimate.

Figure 2.6 displays the number of genes that are differentially expressed (FDR $p < .001$) in a paired test of the control and treatment group for each gene in each strain. As with the analysis of the variance ratios, the spleen has significantly fewer genes differentially expressed than does the skin tissue. Interestingly, Bl6Jx (7,539 genes) and B6.C (7,500 genes) have a very large number of genes differentially expressed in the skin tissue. Differential expression was also calculated between the different strains in each tissue and treatment group. *Figure 2.7* shows the number of genes significantly differentially expressed ($p < .01$ without multiple comparison adjustment) in each paired comparison in the control groups. *Figure 2.8* has the same setup, with the treatment groups. In the control group, there are anywhere from 0 to 12 genes found to be differentially expressed between strains in each tissue. This range drops to be from 0 to 2 in the treatment group. There still seem to be some differences between the tissues, although with such a small number of genes being significant, it is not possible to draw any reliable conclusions. In the skin tissue, there are 1,563 genes found to be

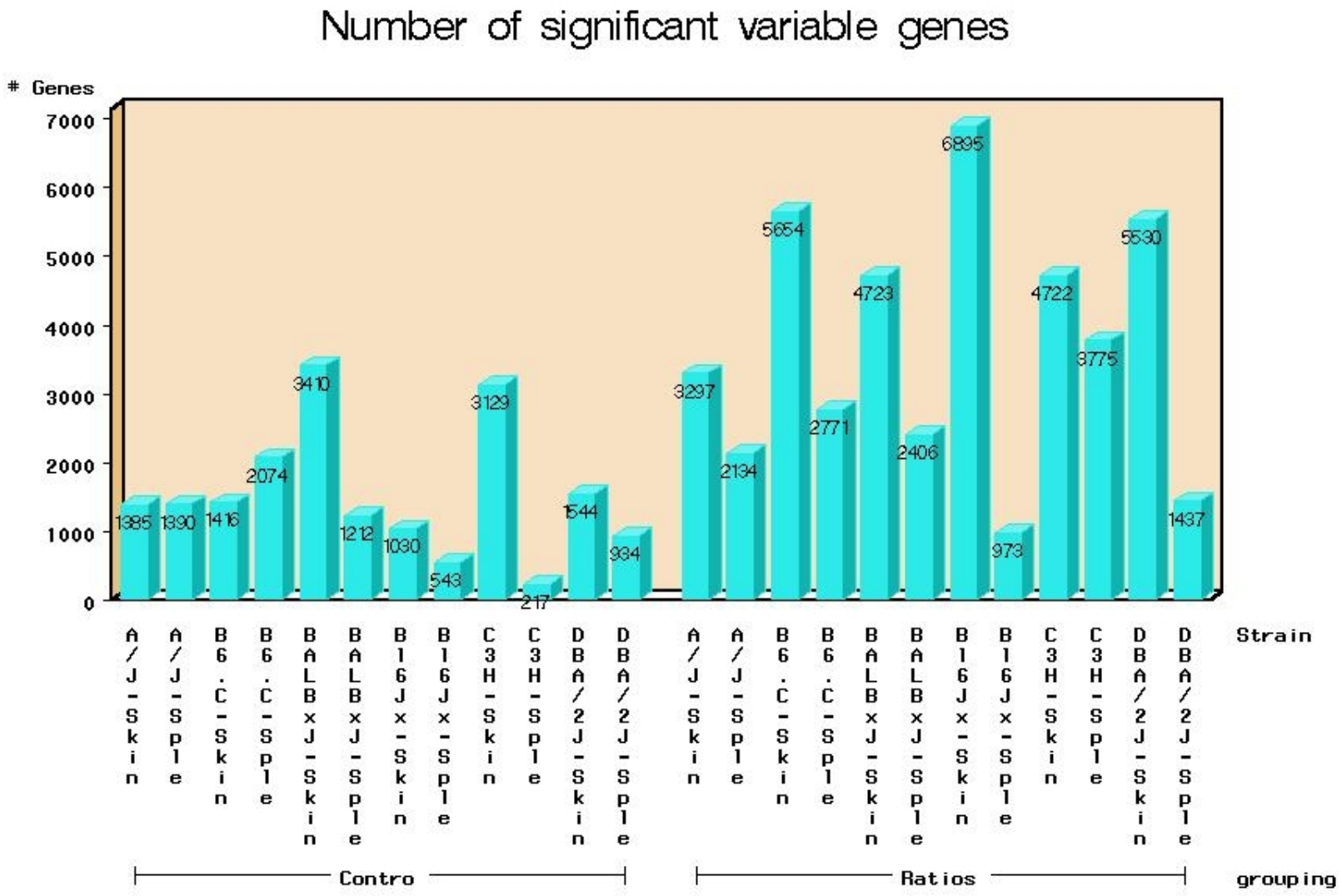


Figure 2.3: Number of genes in each strain and tissue with a statistically significant variance ratio ($p < .05$).

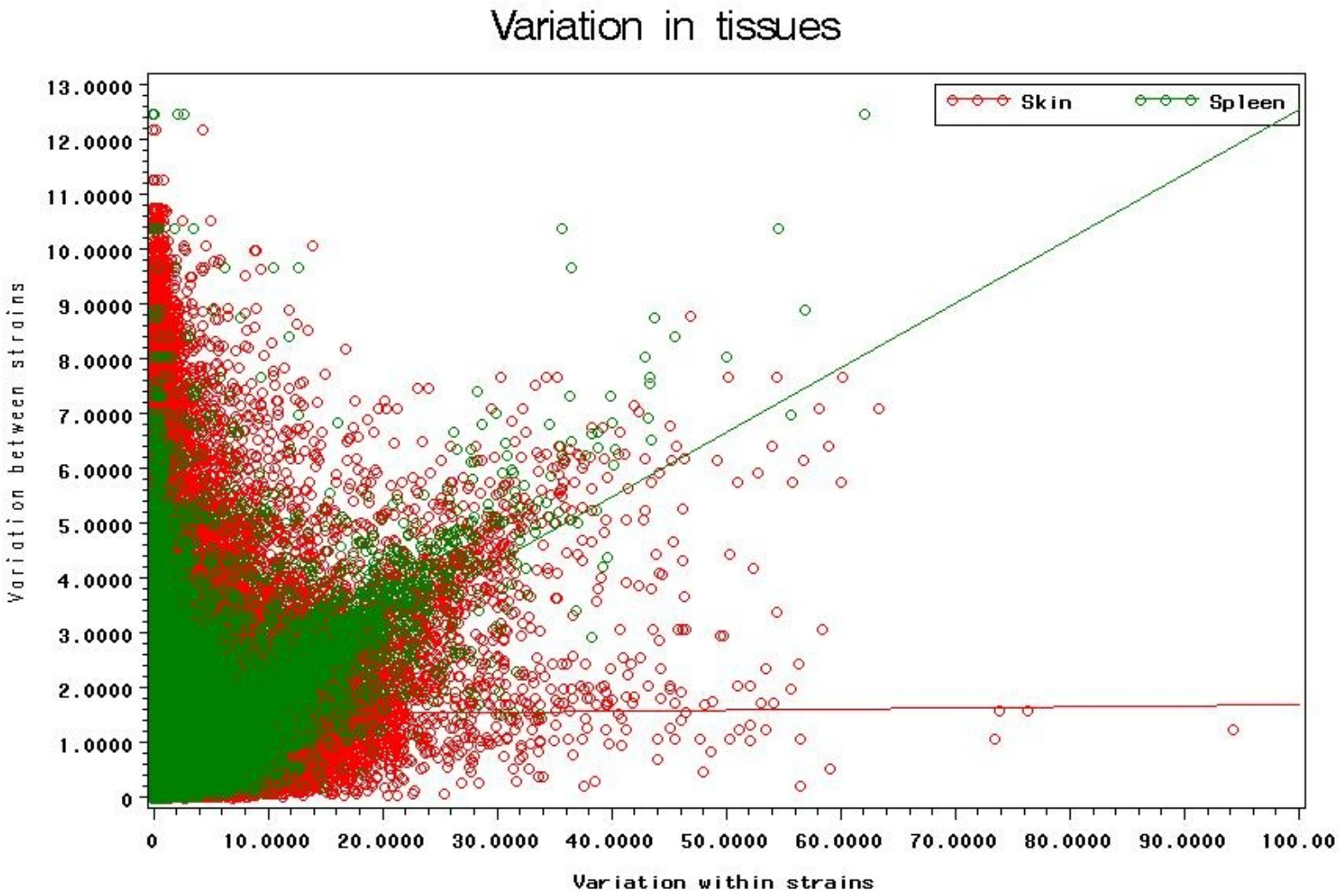


Figure 2.4: REML estimate between strains vs. within strain for the control groups. Red points represent skin tissue and green points represent spleen tissue.

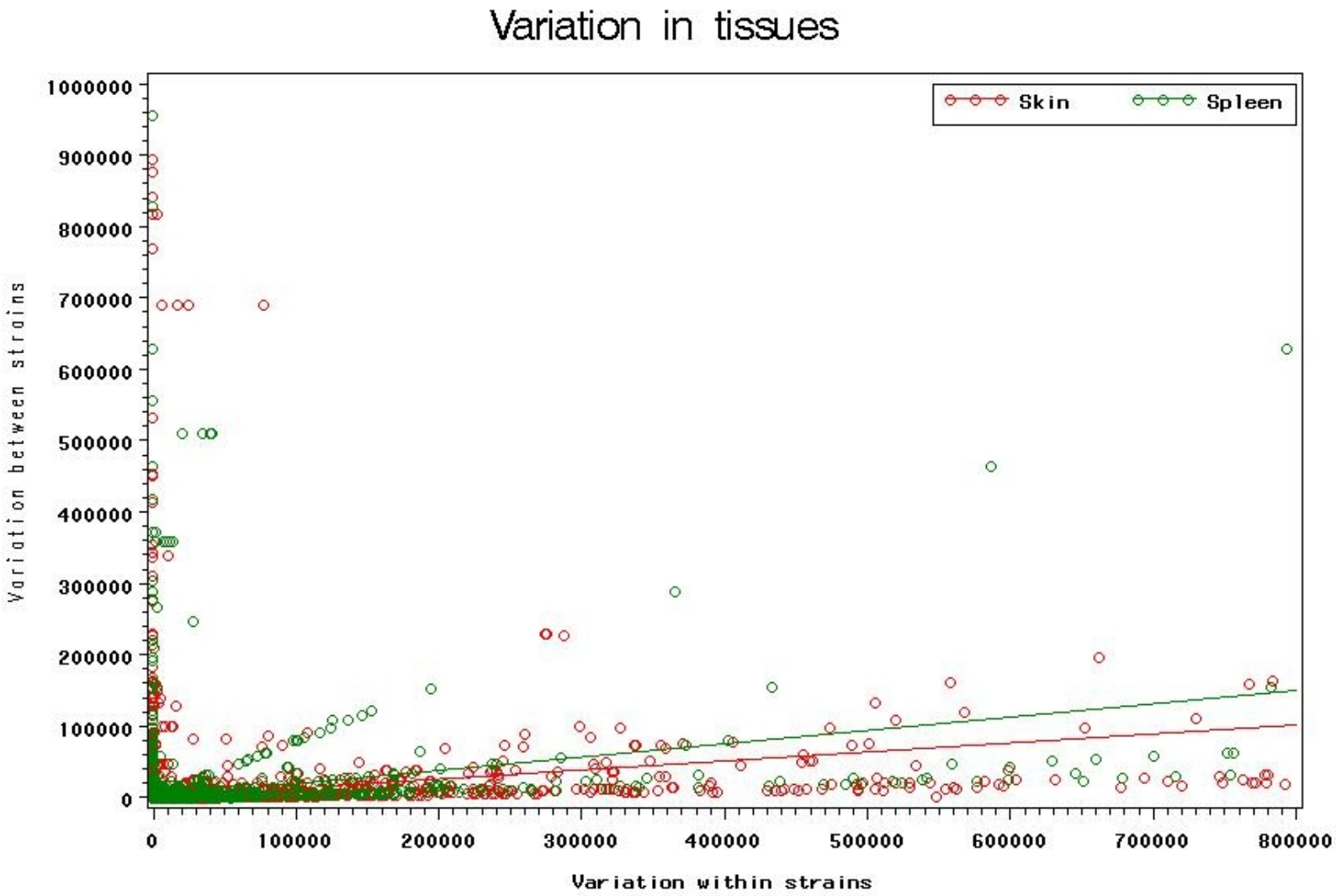


Figure 2.5: REML estimate between strains vs. within strain for the treatment groups. Red points represent skin tissue and green points represent spleen tissue. Several extreme outliers were removed from this plot.

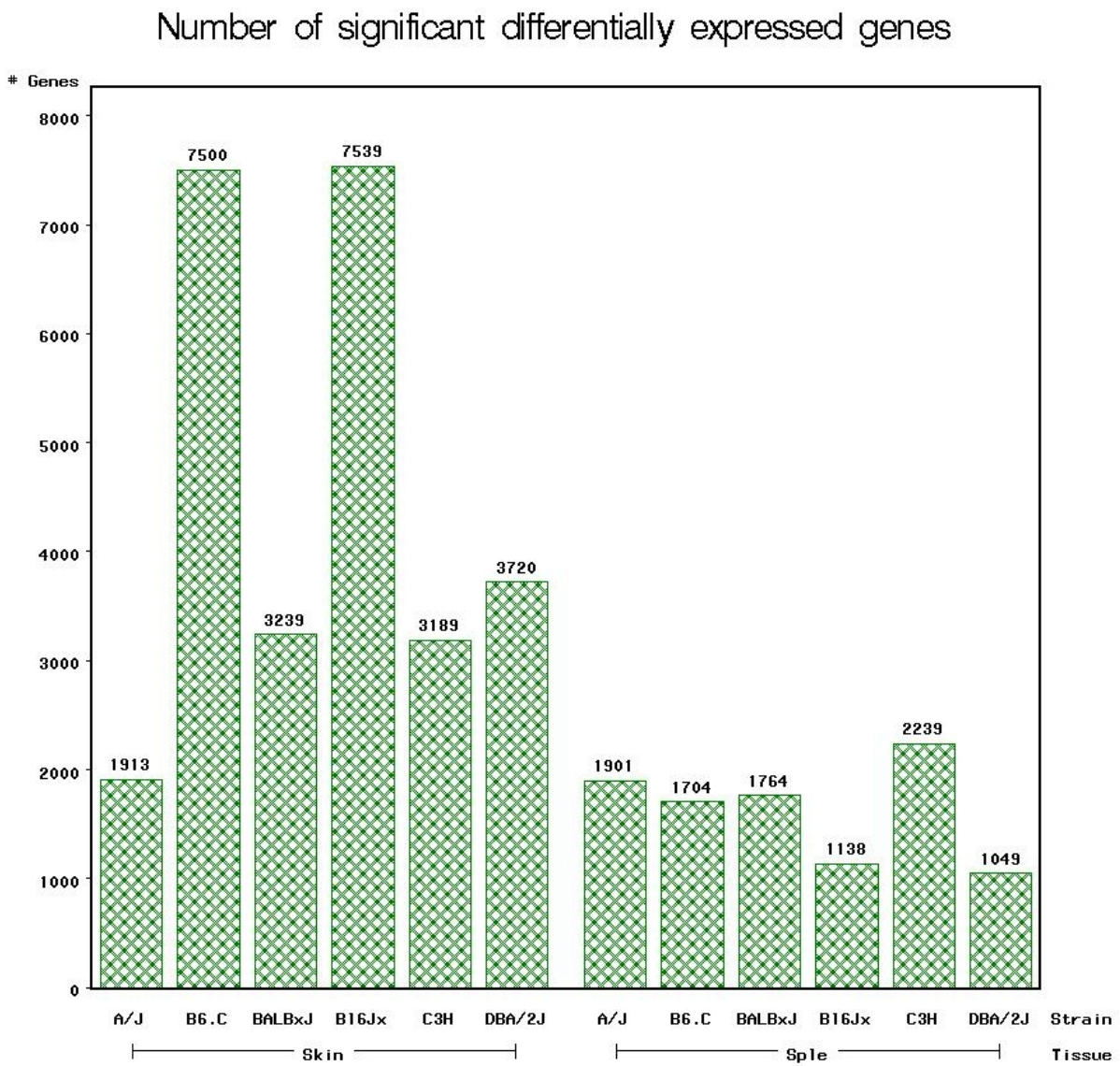


Figure 2.6: The number of significantly differentially expressed genes between the control and treatment groups in each strain and tissue.

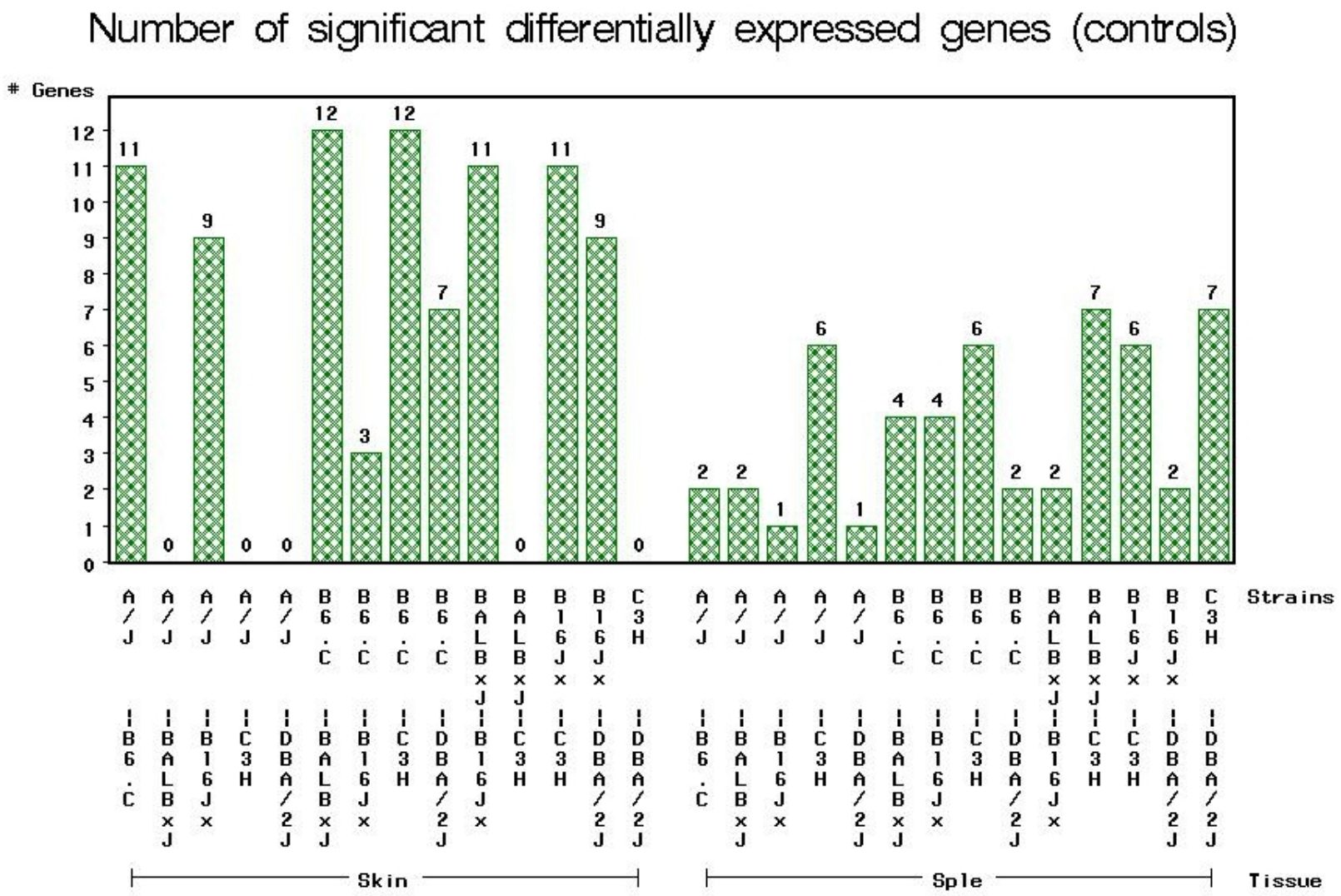


Figure 2.7: The number of significant differentially expressed genes between strains for each tissue in the controls group.

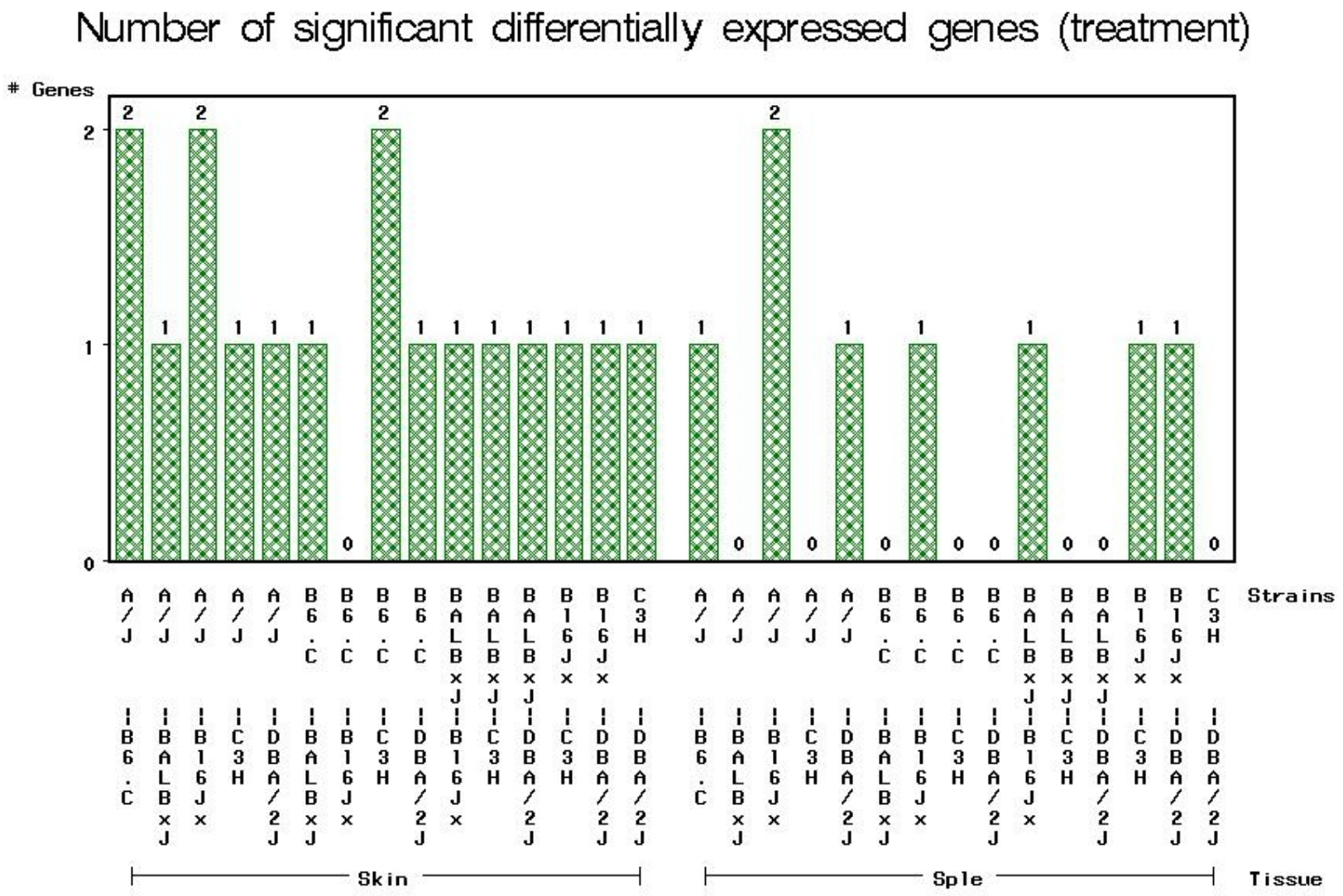


Figure 2.8: The number of significant differentially expressed genes between strains for each tissue in the treatment group.

differentially expressed in all 6 inbred mice strains, and in the spleen tissue, there were 654 genes found.

While the above figures show the number of genes that were differentially expressed, and variable, one question of interest that arises is how many of the genes found to be significant in these different statistical tests are common between genetically distinct organisms, and how many are unique? Additionally, how many genes are common between tissues? These issues are addressed through the subsequent figures and tables. *Table 2.3* displays the number of genes that have a significant variance ratio ($p < .01$) in the control and treatment groups, between the different combinations of strains. All possible combinations of strains were analyzed, and the only group that had genes significantly variable in all six strains was the skin controls (10 genes). *Table 2.4* contains the number of genes that had a significant variance ratio in both the skin and spleen tissues for the control and treatment groups. Very few genes were found to be in common between strains and tissues in this test, and no genes were found to be in common between both tissues and all strains in either the control or treatment groups.

Figure 2.9 displays the number of genes that had a significant variance ratio in one and only one strain in either tissue or control or treatment groups. With the exception of BALBxJ in the skin, and B6.C and Bl6Jx in the spleen, every strain shows an increase in the number of variable genes between the control and treatment groups. The number of genes differentially expressed in one and only one strain in the paired comparison of control and treatment groups (*Figure 2.10*) was significantly less than the number of genes shown in *Figure 2.9*. While the A/J strain was consistently similar in the skin and spleen, the other strains had very few genes differentially expressed in the spleen tissue. In combination of these two groups, as shown in *Figure 2.11*, there were several genes significantly variable in both the control and treatment groups and differentially expressed in a paired comparison between the two in one and only one strain in the skin tissue. There were no such genes that met these criteria in the spleen tissue.

Tests of differential expression between strains yielded very few significant genes, as is shown in *Figures 2.7* and 2.*8*. Even though there are so few genes, a test was done to identify differentially expressed genes common to more than one strain. These results are displayed in *Table 2.5*. There were a few genes found to be in common between 2 paired comparisons between strains in the skin and spleen controls, but none in the treatment groups. *Table 2.6* is a complementing table, showing the number of genes in common between strains in the test for differential expression between the control and treatment groups.

Figures 2.12 and 2.13 show dendrograms produced from the hierarchical clustering of arrays. In *Figure 2.12*, each array is included in the clustering, but the strains do not seem to consistently cluster together, as would be expected. In *Figure 2.13*, the different arrays for a strain are averaged together, to show the relationship between strains. These do not seem to agree between the two tissues in either the control or treatment groups.

Table 2.3: The number of genes with a significant variance ratio that are overlapping between strains.

C-Skin	C-Spleen	T-Skin	T-Spleen	Strain Combination
3410	1212	4723	2406	BALBxJ
1030	543	6895	973	Bl6Jx
1416	2074	5654	2771	B6.C
1544	934	5530	1437	DBA/2J
1385	1390	3297	2134	A/J
3129	217	4722	3775	C3H
525	105	2213	432	BALBxJ and Bl6Jx
699	460	1956	1093	BALBxJ and B6.C
745	144	2286	694	BALBxJ and DBA/2J
635	316	1686	879	BALBxJ and A/J
1625	23	2487	1241	BALBxJ and C3H
433	134	3349	439	Bl6Jx and B6.C
273	56	2546	272	Bl6Jx and DBA/2J
183	90	1608	379	Bl6Jx and A/J
541	10	2340	518	Bl6Jx and C3H
360	265	2183	683	B6.C and DBA/2J
272	548	1420	1075	B6.C and A/J
734	28	2036	1525	B6.C and C3H
269	171	1712	569	DBA/2J and A/J
666	14	2256	729	DBA/2J and C3H
562	23	1681	1164	A/J and C3H
219	41	1135	229	BALBxJ, Bl6Jx and B6.C
146	12	1135	166	BALBxJ, Bl6Jx and DBA/2J
105	35	832	204	BALBxJ, Bl6Jx and A/J
335	0	1229	239	BALBxJ, Bl6Jx and C3H
196	76	980	379	BALBxJ, B6.C and DBA/2J
159	167	722	485	BALBxJ, B6.C and A/J
427	5	1030	615	BALBxJ, B6.C and C3H
161	44	958	334	BALBxJ, DBA/2J and A/J
418	4	1317	388	BALBxJ, DBA/2J and C3H
374	2	1046	490	BALBxJ, A/J and C3H
111	19	1313	158	Bl6Jx, B6.C and DBA/2J
91	40	804	221	Bl6Jx, B6.C and A/J
260	0	1261	247	Bl6Jx, B6.C and C3H
43	14	876	145	Bl6Jx, DBA/2J and A/J
149	2	1175	142	Bl6Jx, DBA/2J and C3H
95	2	846	218	Bl6Jx, A/J and C3H
63	86	752	336	B6.C, DBA/2J and A/J
198	1	988	394	B6.C, DBA/2J and C3H
161	6	716	647	B6.C, A/J and C3H
146	3	956	311	DBA/2J, A/J and C3H
54	7	528	95	BALBxJ, Bl6Jx, B6.C and DBA/2J
56	17	363	114	BALBxJ, Bl6Jx, B6.C and A/J
148	0	572	106	BALBxJ, Bl6Jx, B6.C and C3H
23	5	445	89	BALBxJ, Bl6Jx, DBA/2J and A/J
92	0	635	74	BALBxJ, Bl6Jx, DBA/2J and C3H
67	0	490	104	BALBxJ, Bl6Jx, A/J and C3H
39	29	357	192	BALBxJ, B6.C, DBA/2J and A/J
117	0	500	208	BALBxJ, B6.C, DBA/2J and C3H
111	1	397	259	BALBxJ, B6.C, A/J and C3H
100	1	581	165	BALBxJ, DBA/2J, A/J and C3H
21	9	382	91	Bl6Jx, B6.C, DBA/2J and A/J
73	0	575	71	Bl6Jx, B6.C, DBA/2J and C3H
56	0	372	114	Bl6Jx, B6.C, A/J and C3H
26	0	451	61	Bl6Jx, DBA/2J, A/J and C3H
39	1	354	177	B6.C, DBA/2J, A/J and C3H
12	5	126	52	BALBxJ, Bl6Jx, B6.C, DBA/2J and A/J
38	0	230	28	BALBxJ, Bl6Jx, B6.C, DBA/2J and C3H
40	0	166	38	BALBxJ, Bl6Jx, B6.C, A/J and C3H
15	0	225	24	BALBxJ, Bl6Jx, DBA/2J, A/J and C3H
25	0	153	80	BALBxJ, B6.C, DBA/2J, A/J and C3H
16	0	131	26	Bl6Jx, B6.C, DBA/2J, A/J and C3H
10	0	0	0	BALBxJ, Bl6Jx, B6.C, DBA/2J, A/J and C3H

Table 2.4: The number of genes with a significant variance ratio in the skin and spleen that are overlapping between strains.

Control	Treatment	Strain Combination
115	544	BALBxJ
10	297	Bl6Jx
32	730	B6.C
24	403	DBA/2J
57	381	A/J
17	867	C3H
6	56	BALBxJ and Bl6Jx
16	98	BALBxJ and B6.C
0	36	BALBxJ and DBA/2J
5	73	BALBxJ and A/J
0	111	BALBxJ and C3H
0	75	Bl6Jx and B6.C
0	78	Bl6Jx and DBA/2J
0	35	Bl6Jx and A/J
0	57	Bl6Jx and C3H
0	43	B6.C and DBA/2J
2	106	B6.C and A/J
0	356	B6.C and C3H
6	23	DBA/2J and A/J
0	41	DBA/2J and C3H
1	205	A/J and C3H
2	11	BALBxJ, Bl6Jx and B6.C
0	56	BALBxJ, Bl6Jx and DBA/2J
1	9	BALBxJ, Bl6Jx and A/J
0	4	BALBxJ, Bl6Jx and C3H
0	24	BALBxJ, B6.C and DBA/2J
3	18	BALBxJ, B6.C and A/J
0	55	BALBxJ, B6.C and C3H
0	14	BALBxJ, DBA/2J and A/J
0	18	BALBxJ, DBA/2J and C3H
0	37	BALBxJ, A/J and C3H
0	18	Bl6Jx, B6.C and DBA/2J
0	13	Bl6Jx, B6.C and A/J
0	29	Bl6Jx, B6.C and C3H
0	17	Bl6Jx, DBA/2J and A/J
0	12	Bl6Jx, DBA/2J and C3H
0	13	Bl6Jx, A/J and C3H
0	5	B6.C, DBA/2J and A/J
0	19	B6.C, DBA/2J and C3H
0	106	B6.C, A/J and C3H
0	8	DBA/2J, A/J and C3H
0	11	BALBxJ, Bl6Jx, B6.C and DBA/2J
1	0	BALBxJ, Bl6Jx, B6.C and A/J
0	0	BALBxJ, Bl6Jx, B6.C and C3H
0	9	BALBxJ, Bl6Jx, DBA/2J and A/J
0	4	BALBxJ, Bl6Jx, DBA/2J and C3H
0	0	BALBxJ, Bl6Jx, A/J and C3H
0	0	BALBxJ, B6.C, DBA/2J and A/J
0	5	BALBxJ, B6.C, DBA/2J and C3H
0	18	BALBxJ, B6.C, A/J and C3H
0	2	BALBxJ, DBA/2J, A/J and C3H
0	5	Bl6Jx, B6.C, DBA/2J and A/J
0	2	Bl6Jx, B6.C, DBA/2J and C3H
0	13	Bl6Jx, B6.C, A/J and C3H
0	5	Bl6Jx, DBA/2J, A/J and C3H
0	5	B6.C, DBA/2J, A/J and C3H
0	0	BALBxJ, Bl6Jx, B6.C, DBA/2J and A/J
0	0	BALBxJ, Bl6Jx, B6.C, DBA/2J and C3H
0	0	BALBxJ, Bl6Jx, B6.C, A/J and C3H
0	0	BALBxJ, Bl6Jx, DBA/2J, A/J and C3H
0	0	BALBxJ, B6.C, DBA/2J, A/J and C3H
0	0	Bl6Jx, B6.C, DBA/2J, A/J and C3H
0	0	BALBxJ, Bl6Jx, B6.C, DBA/2J, A/J and C3H

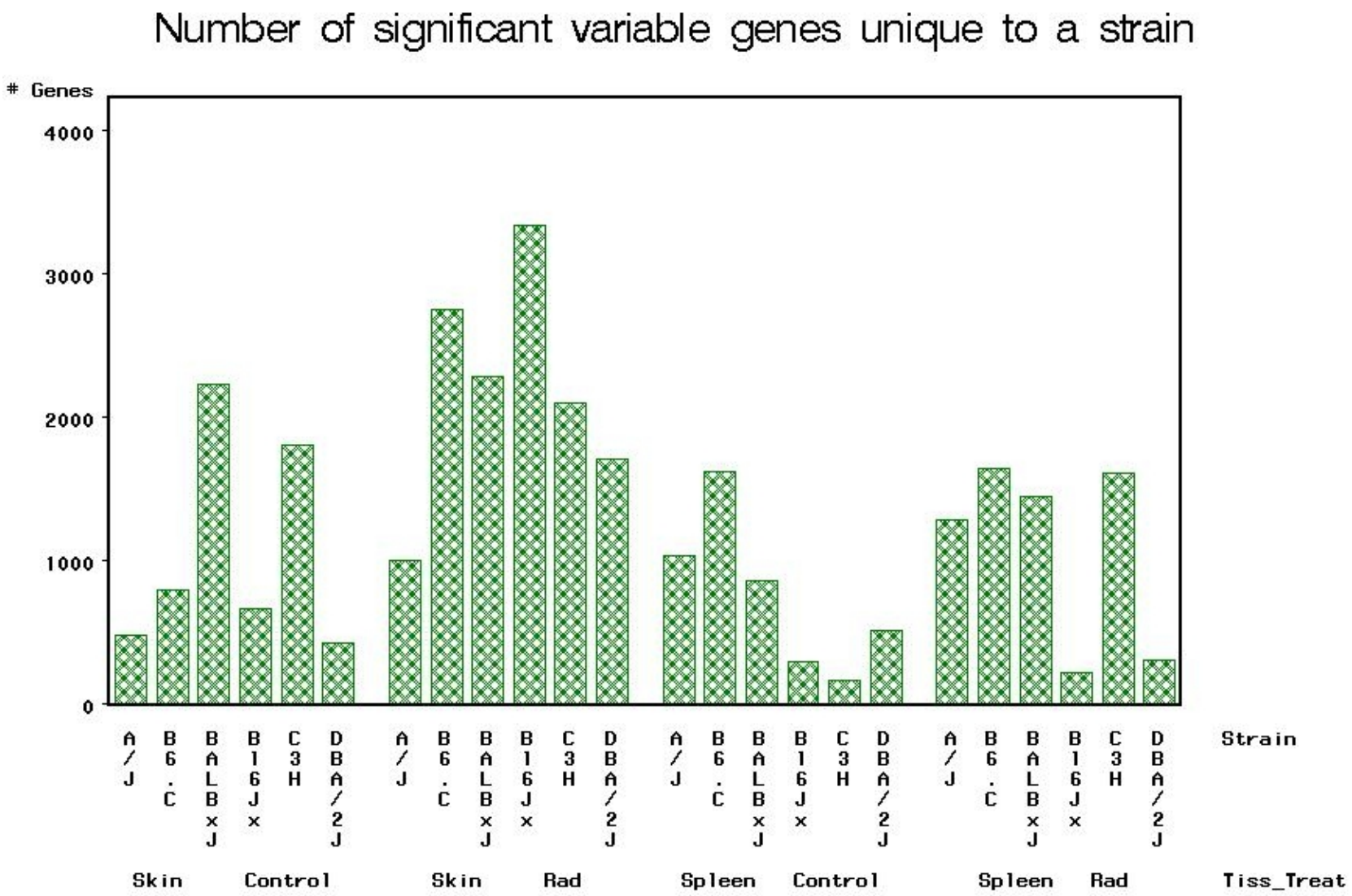


Figure 2.9: The number of genes with a significant variance ratio that were unique to a particular strain within each group.

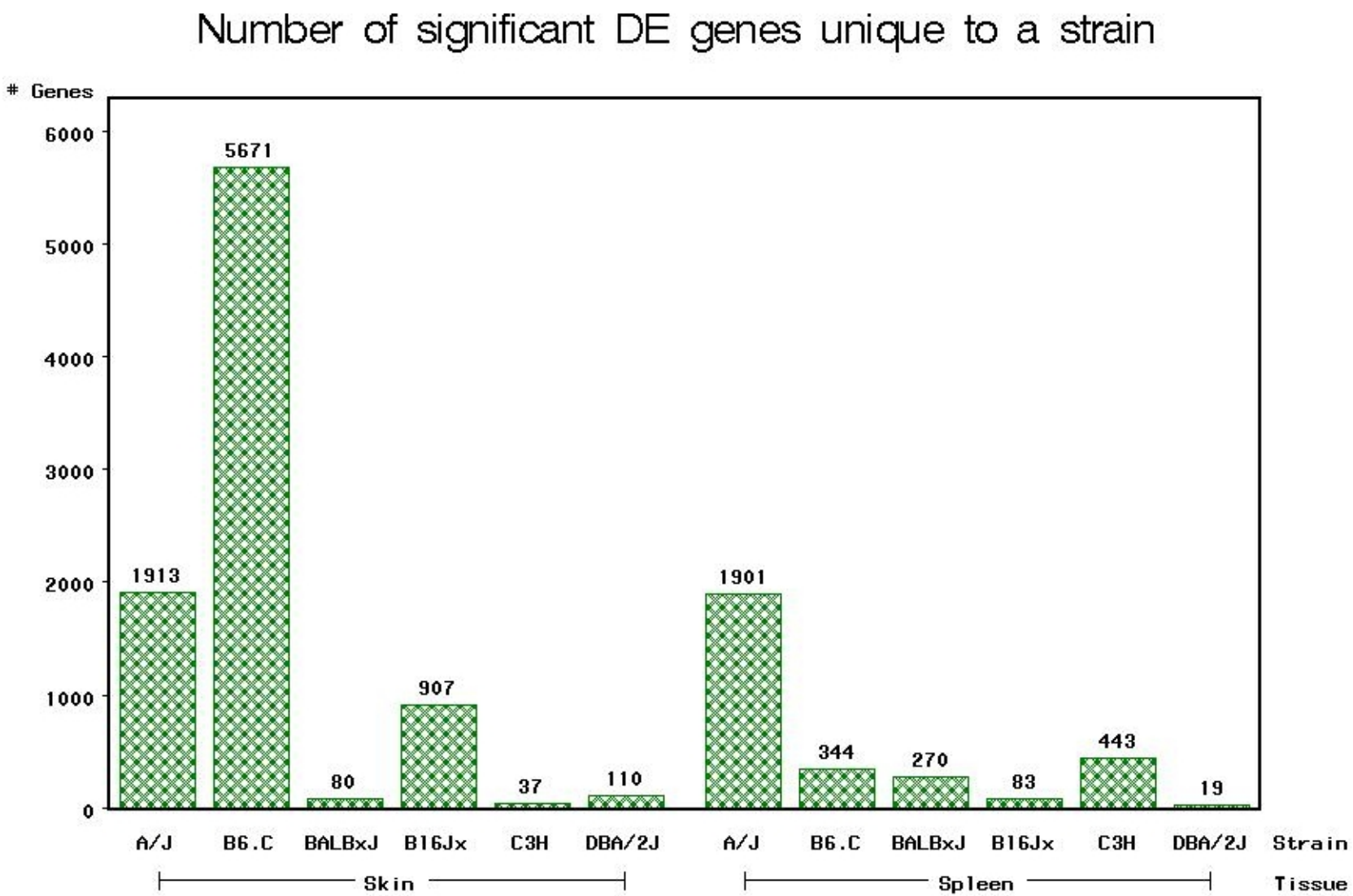


Figure 2.10: The number of genes significantly differentially expressed between the control and treatment groups, which were unique to a particular strain in each of the two tissues.

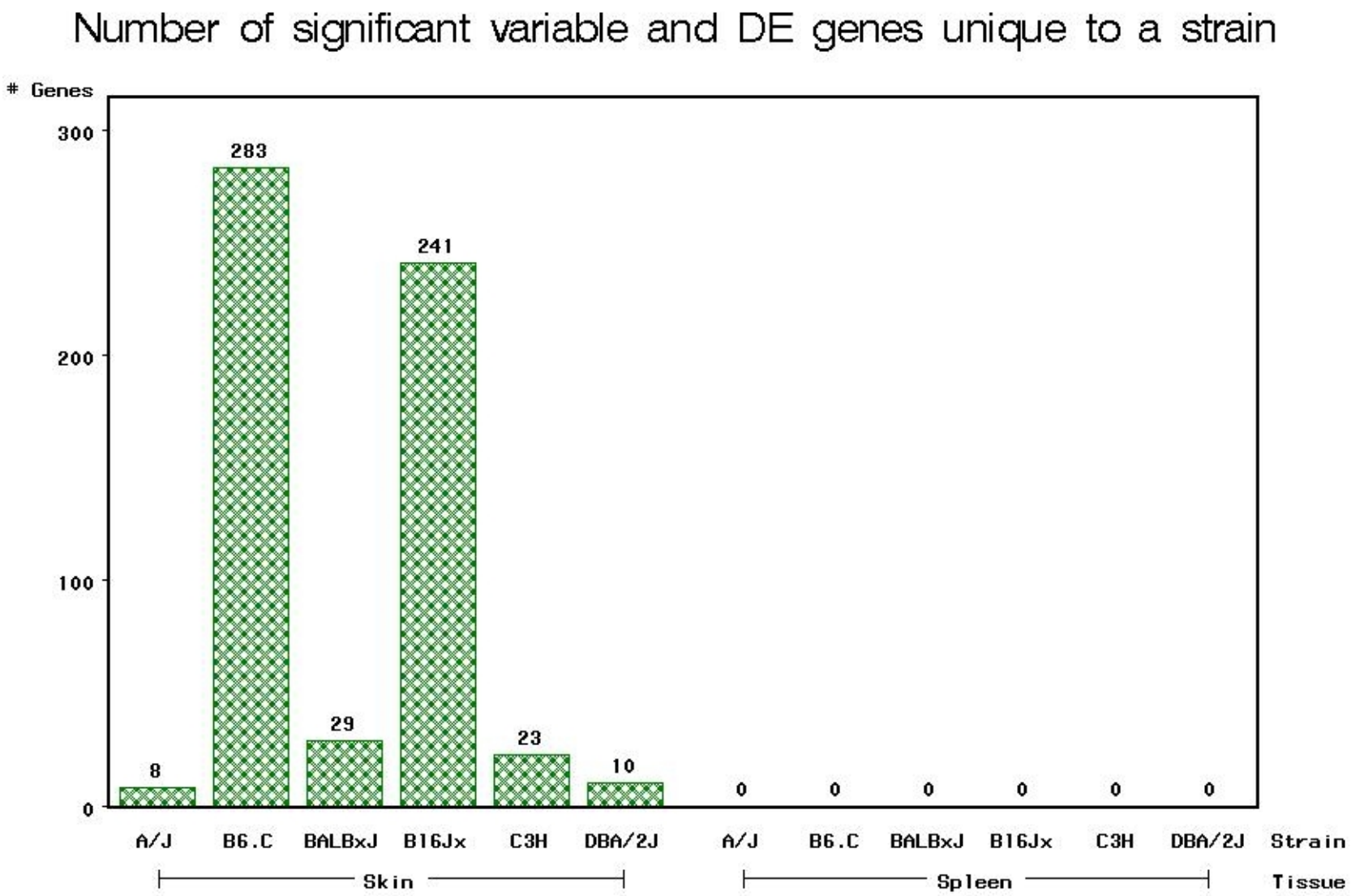


Figure 2.11: The number of genes significantly differentially expressed between the control and treatment groups and also had a significant variance ratio, that were unique to a particular strain in each of the two tissues.

Table 2.5: The number of genes differentially expressed between strains.

C-Skin	C-Spleen	T-Skin	T-Spleen	Strain Comparison
11	2	2	1	A/J and B6.C
0	2	1	0	A/J and BALBxJ
9	1	2	2	A/J and Bl6Jx
0	6	1	0	A/J and C3H
0	1	1	1	A/J and DBA/2J
12	4	1	0	B6.C and BALBxJ
3	4	0	1	B6.C and Bl6Jx
12	6	2	0	B6.C and C3H
7	2	1	0	B6.C and DBA/2J
11	2	1	1	BALBxJ and Bl6Jx
0	7	1	0	BALBxJ and C3H
0	0	1	0	BALBxJ and DBA/2J
11	6	1	1	Bl6Jx and C3H
9	2	1	1	Bl6Jx and DBA/2J
0	7	1	0	C3H and DBA/2J
0	1	0	0	A/J, B6.C and BALBxJ
1	0	0	0	A/J, B6.C and Bl6Jx
0	1	0	0	A/J, B6.C and C3H
0	0	0	0	A/J, B6.C and DBA/2J
0	0	0	0	A/J, BALBxJ and Bl6Jx
0	2	0	0	A/J, BALBxJ and C3H
0	0	0	0	A/J, BALBxJ and DBA/2J
0	1	0	0	A/J, Bl6Jx and C3H
0	0	0	0	A/J, Bl6Jx and DBA/2J
0	1	0	0	A/J, C3H and DBA/2J
2	0	0	0	B6.C, BALBxJ and Bl6Jx
0	3	0	0	B6.C, BALBxJ and C3H
0	0	0	0	B6.C, BALBxJ and DBA/2J
2	3	0	0	B6.C, Bl6Jx and C3H
1	0	0	0	B6.C, Bl6Jx and DBA/2J
0	2	0	0	B6.C, C3H and DBA/2J
0	2	0	0	BALBxJ, Bl6Jx and C3H
0	0	0	0	BALBxJ, Bl6Jx and DBA/2J
0	0	0	0	BALBxJ, C3H and DBA/2J
0	1	0	0	Bl6Jx, C3H and DBA/2J

Table 2.6: The number of genes differentially expressed between the control and treatment groups.

Skin	Spleen	Strain Combination
3239	1764	BALBxJ
7539	1138	Bl6Jx
7500	1704	B6.C
3720	1049	DBA/2J
1913	1901	A/J
3189	2239	C3H
3105	987	BALBxJ and Bl6Jx
3150	1295	BALBxJ and B6.C
2626	972	BALBxJ and DBA/2J
1700	1366	BALBxJ and A/J
2667	1344	BALBxJ and C3H
6614	913	Bl6Jx and B6.C
3491	758	Bl6Jx and DBA/2J
1818	960	Bl6Jx and A/J
3086	926	Bl6Jx and C3H
3546	922	B6.C and DBA/2J
1829	1360	B6.C and A/J
3114	1386	B6.C and C3H
1704	934	DBA/2J and A/J
2615	876	DBA/2J and C3H
1668	1437	A/J and C3H
3093	875	BALBxJ, Bl6Jx and B6.C
2580	751	BALBxJ, Bl6Jx and DBA/2J
1686	907	BALBxJ, Bl6Jx and A/J
2623	856	BALBxJ, Bl6Jx and C3H
2608	881	BALBxJ, B6.C and DBA/2J
1691	1167	BALBxJ, B6.C and A/J
2645	1119	BALBxJ, B6.C and C3H
1628	892	BALBxJ, DBA/2J and A/J
2347	833	BALBxJ, DBA/2J and C3H
1626	1135	BALBxJ, A/J and C3H
3451	722	Bl6Jx, B6.C and DBA/2J
1812	863	Bl6Jx, B6.C and A/J
3068	815	Bl6Jx, B6.C and C3H
1691	732	Bl6Jx, DBA/2J and A/J
2564	682	Bl6Jx, DBA/2J and C3H
1659	837	Bl6Jx, A/J and C3H
1697	873	B6.C, DBA/2J and A/J
2592	825	B6.C, DBA/2J and C3H
1662	1171	B6.C, A/J and C3H
1604	816	DBA/2J, A/J and C3H
2578	715	BALBxJ, Bl6Jx, B6.C and DBA/2J
1686	840	BALBxJ, Bl6Jx, B6.C and A/J
2620	790	BALBxJ, Bl6Jx, B6.C and C3H
1617	726	BALBxJ, Bl6Jx, DBA/2J and A/J
2336	676	BALBxJ, Bl6Jx, DBA/2J and C3H
1618	805	BALBxJ, Bl6Jx, A/J and C3H
1621	840	BALBxJ, B6.C, DBA/2J and A/J
2313	794	BALBxJ, B6.C, DBA/2J and C3H
1621	1026	BALBxJ, B6.C, A/J and C3H
1571	786	BALBxJ, DBA/2J, A/J and C3H
1691	708	Bl6Jx, B6.C, DBA/2J and A/J
2559	666	Bl6Jx, B6.C, DBA/2J and C3H
1659	784	Bl6Jx, B6.C, A/J and C3H
1596	669	Bl6Jx, DBA/2J, A/J and C3H
1599	791	B6.C, DBA/2J, A/J and C3H
1617	702	BALBxJ, Bl6Jx, B6.C, DBA/2J and A/J
2312	660	BALBxJ, Bl6Jx, B6.C, DBA/2J and C3H
1618	769	BALBxJ, Bl6Jx, B6.C, A/J and C3H
1563	664	BALBxJ, Bl6Jx, DBA/2J, A/J and C3H
1566	765	BALBxJ, B6.C, DBA/2J, A/J and C3H
1596	659	Bl6Jx, B6.C, DBA/2J, A/J and C3H
1563	654	BALBxJ, Bl6Jx, B6.C, DBA/2J, A/J and C3H

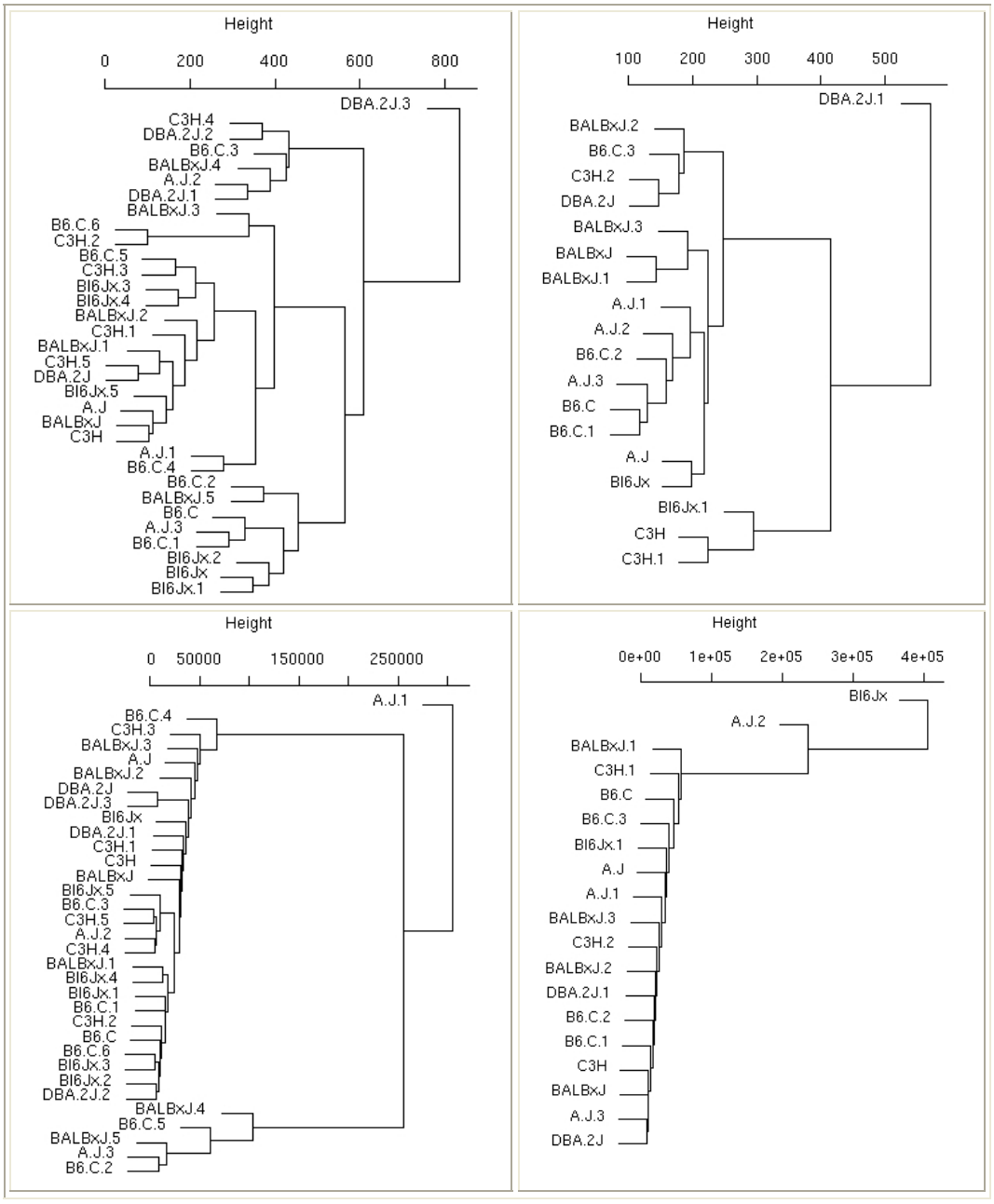


Figure 2.12: Hierarchical clustering of arrays. Top Left: Skin Controls, Top Right: Spleen Controls, Bottom Left: Skin Treatment, Bottom Right: Spleen Treatment.

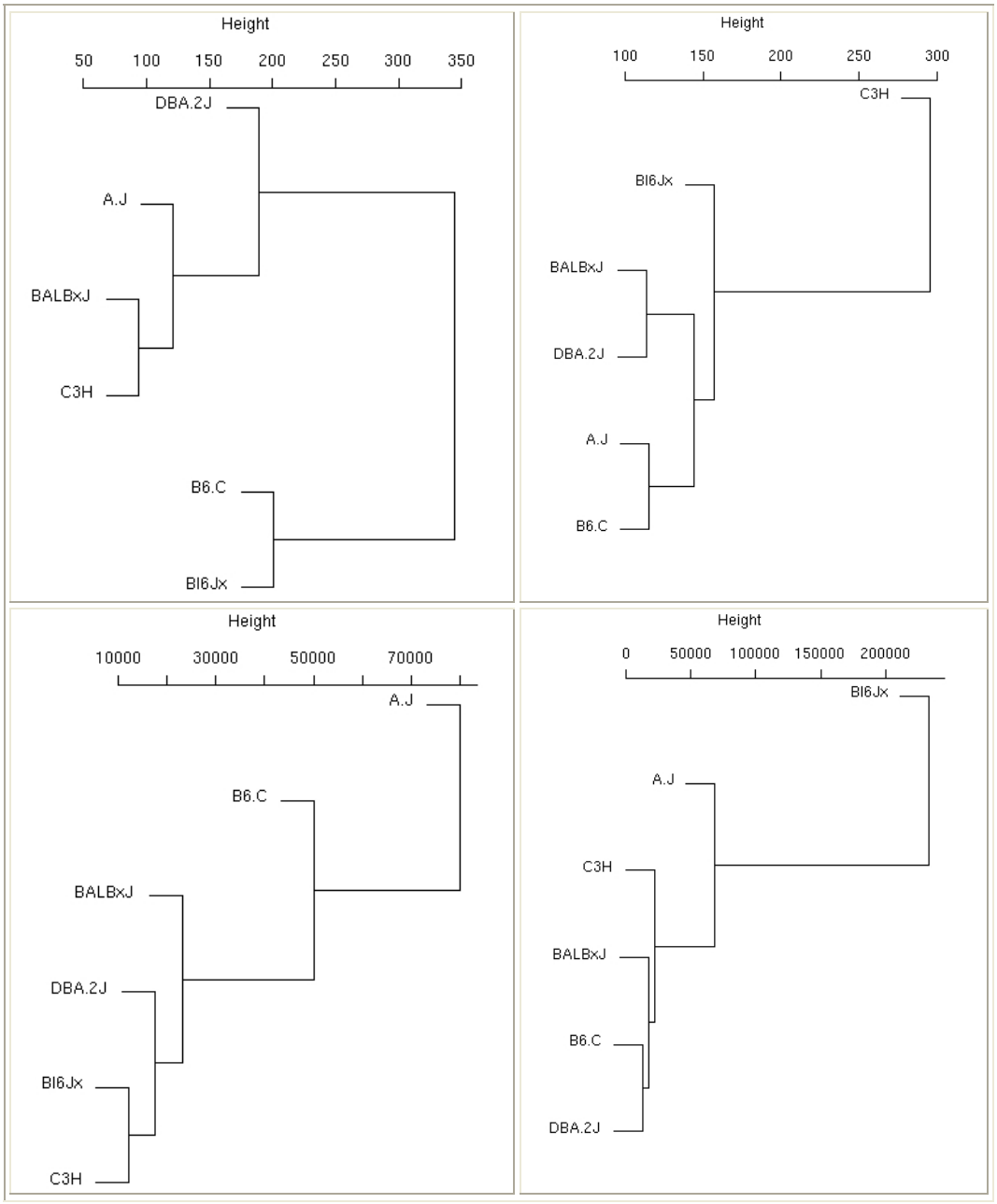


Figure 2.13: Hierarchical clustering of strains with arrays averaged. Top Left: Skin Controls, Top Right: Spleen Controls, Bottom Left: Skin Treatment, Bottom Right: Spleen Treatment.

Gene Set Analysis:

There were many different combinations of the above results created, and subsequently analyzed for significance. The results of the analysis of each set are detailed in the following section.

1. Genes with a significant variance ratio and were common between 6 of 6 strains in the skin control. There were 10 genes in this set, all of which had a Gene ID. 4 of these 10 genes did not have a chromosomal location assigned. There were no chromosomes with three or more genes, so density plots were not created. There were 7 genes that did not have Gene Ontology annotations, and from the remaining 3 genes in the set, no Gene Ontology categories were enriched.

2. Genes with a significant variance ratio and were common between 5 of 6 strains in the spleen control. There were 5 genes in this set, all of which had a Gene ID. There were no chromosomes with three or more genes, so density plots were not created. There were 2 genes that did not have Gene Ontology annotations, and from the remaining 3 genes in the set, no Gene Ontology categories were enriched.

3. Genes with a significant variance ratio in 5 of 6 strains in the skin treatment group. There were 1,031 genes in this set, 173 of which did not have a Gene ID. This reduced the set to 858 genes, 151 of which did not have an assigned chromosomal location. The gene density plots of each chromosome are in *Figure 2.15*. There were no gene-dense regions found on any of the chromosomes from this set.

Since this set was so large, clustering was performed using the HOPACH algorithm. There were three variables used for clustering: the average variance ratio for the 5 strains, the maximum variance ratio and the range of the variance ratios. From this data, the HOPACH algorithm identified 190 main clusters in the 5^{th} level of the tree, with a median split silhouette of 0.2848881. *Figure 2.14* shows the ordered distance matrix of the genes in the set. The overwhelming amount of red in the plot signifies that there is very little dissimilarity between the genes in the set, and between clusters. There are some genes that apparently have no correlation as indicated by the yellow regions. *Figure 2.14* also shows the barplot of the bootstrap reappearance proportions. Most genes belong to the first cluster, as indicated by the red vertical bar on the left-hand side of the plot. Most genes appear in 3 to 4 different clusters, at varying proportions.

Since there were so many clusters derived by HOPACH, clusters with less than 10 genes in them were combined into clusters of 20 or more genes, as described in the Methods section. This yielded 124 clusters that had at least 10 genes each in them. From this set, there were 209 genes that did not have Gene Ontology annotations. Each cluster was analyzed for enriched terms, as described in the Methods section. There were three terms found to be enriched: cytokine and chemokine mediated signaling pathway (biological process), heparin binding (molecular function), and integrin-mediated signaling pathway (biological process).

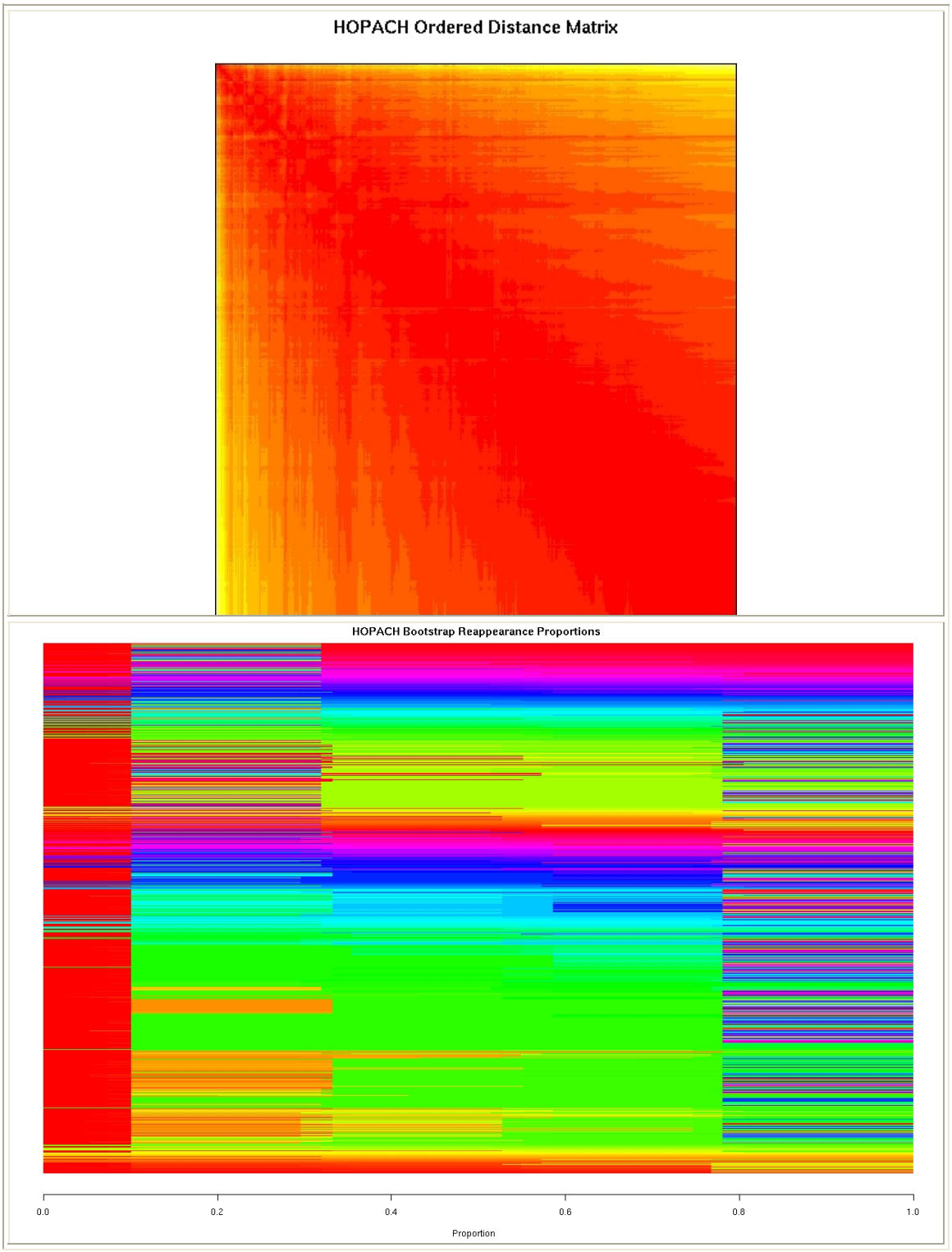


Figure 2.14: Ordered distance matrix and bootstrap reappearance proportions of genes with a significant variance ratio that were common between 5 of 6 strains in the skin treatment group.

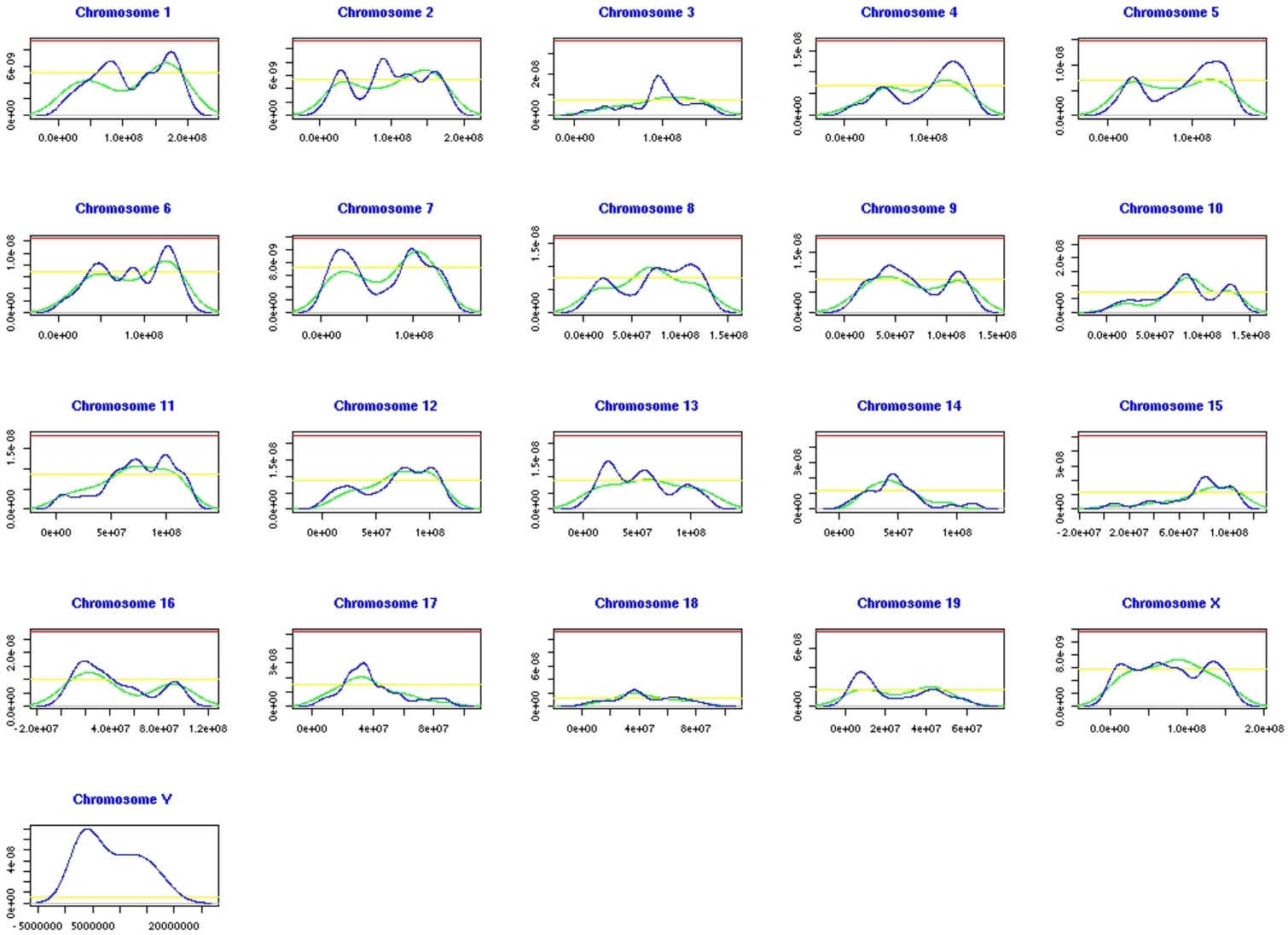


Figure 2.15: Gene density plots of genes with a significant variance ratio and were common between 5 of 6 strains in the skin treatment group.

4. Genes with a significant variance ratio in 5 of 6 strains in the spleen treatment group. There were 248 genes in this set, 24 of which didn't have a Gene ID. This reduced the set to 224 genes, 23 of which did not have an assigned chromosomal location. The gene density plots of each chromosome are in *Figure 2.17*. There were no gene-dense regions found on any of the chromosomes from this set.

Since this set was so large, clustering was performed using the HOPACH algorithm. There were three variables used for clustering: the average variance ratio for the 5 strains, the maximum variance ratio and the range of the variance ratios. From this data, the HOPACH algorithm identified 65 main clusters in the 5^{th} level of the tree, with a median split silhouette of 0.3077117. *Figure 2.16* shows the ordered distance matrix of the genes in this set. Almost all of the genes have a positive correlation (very little dissimilarity), although there are a few that have no correlation. *Figure 2.16* also shows the barplot of the bootstrap reappearance proportions. Very few of the genes in this set belong to only one cluster, although most genes seem to be a large component of one cluster, and small components of several others.

Since there were so many clusters derived by HOPACH, clusters with less than 10 genes in them were combined into clusters of 20 or more genes, as described in the Methods section. This yielded 28 clusters that had at least 10 genes each in them. From this set, there were 22 genes that did not have Gene Ontology annotations. Each cluster was analyzed for enriched terms, as described in the Methods section, but no terms were found to be enriched in this set.

5. Genes with significant variance between the strains, but not within the strain in the skin treatment group (within strain variance < 20,000 and between strain variance > 100,000). There were 55 genes in this set, 7 of which did not have a Gene ID. This reduced the set to 48 genes, 7 of which did not have an assigned chromosomal location. The gene density plots of each chromosome are in *Figure 2.19*. There were no gene-dense regions found on any of the chromosomes from this set.

Since this set was somewhat large, clustering was performed using the HOPACH algorithm. There were three variables used for clustering: the average variance ratio for the strains, the maximum variance ratio and the range of the variance ratios. From this data, the HOPACH algorithm identified 18 main clusters in the 3^{rd} level of the tree, with a median split silhouette of 0.1860477. *Figure 2.18* shows the ordered distance matrix of the genes in this set. Many of the genes have a positive correlation (very little dissimilarity), although there are about equally as many that have no correlation. *Figure 2.18* also shows the barplot of the bootstrap reappearance proportions. About half of the genes in this set belong to only one cluster, and the other half of the genes belong to two or three clusters.

Since there were so many clusters derived by HOPACH, clusters with less than 10 genes in them were combined into clusters of 20 or more genes, as described in the Methods section. This yielded 3 clusters that had at least 10 genes each in them. From this set, there were 14 genes that did not have Gene Ontology annotations. Each cluster was analyzed for enriched terms, as described in the Methods section, but no terms were found to be enriched in this set.



Figure 2.16: Ordered distance matrix and bootstrap reappearance proportions of genes with a significant variance ratio in 5 of 6 strains in the spleen treatment group.

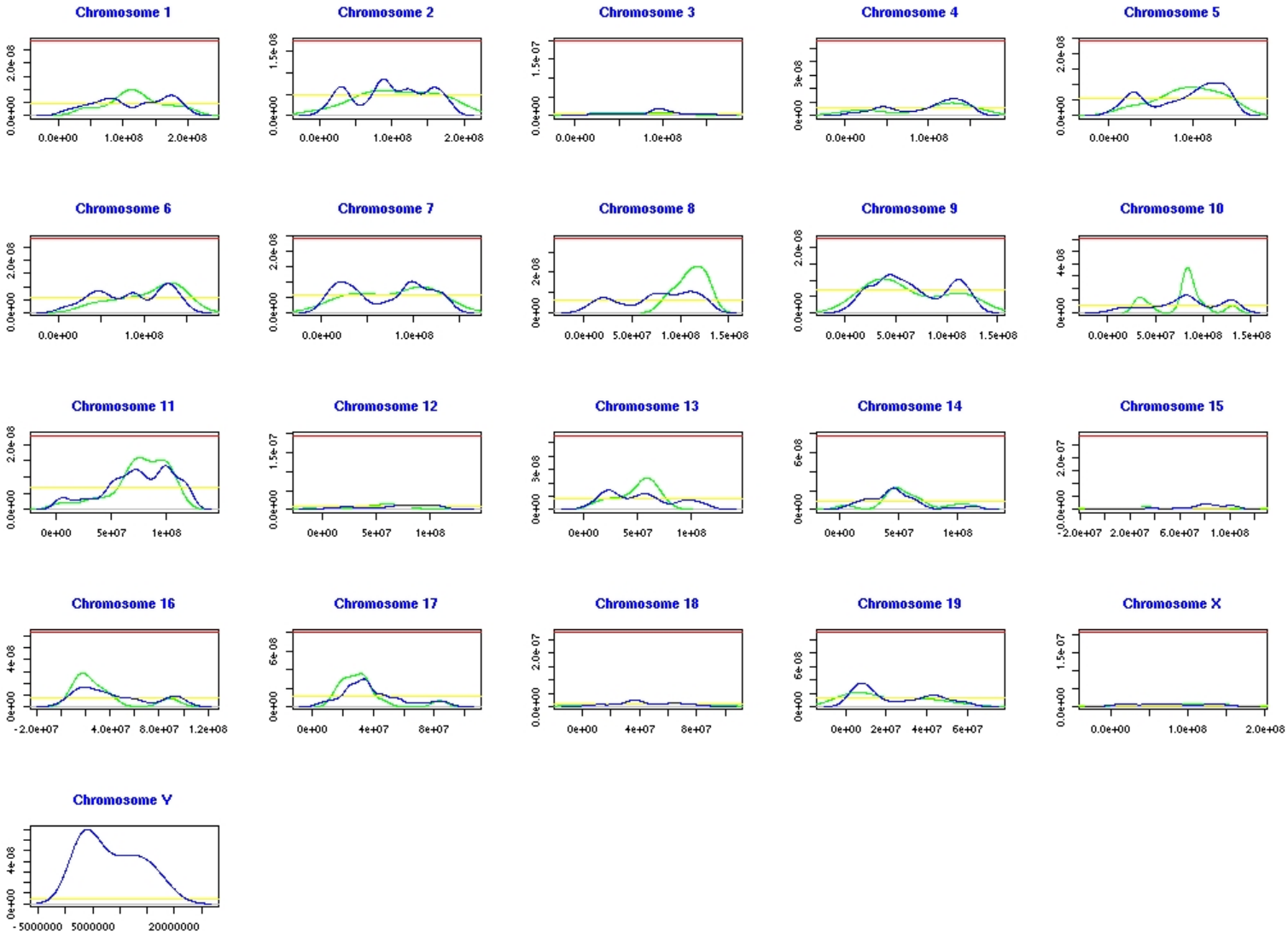


Figure 2.17: Gene density plots of genes with a significant variance ratio in 5 of 6 strains in the spleen treatment group.

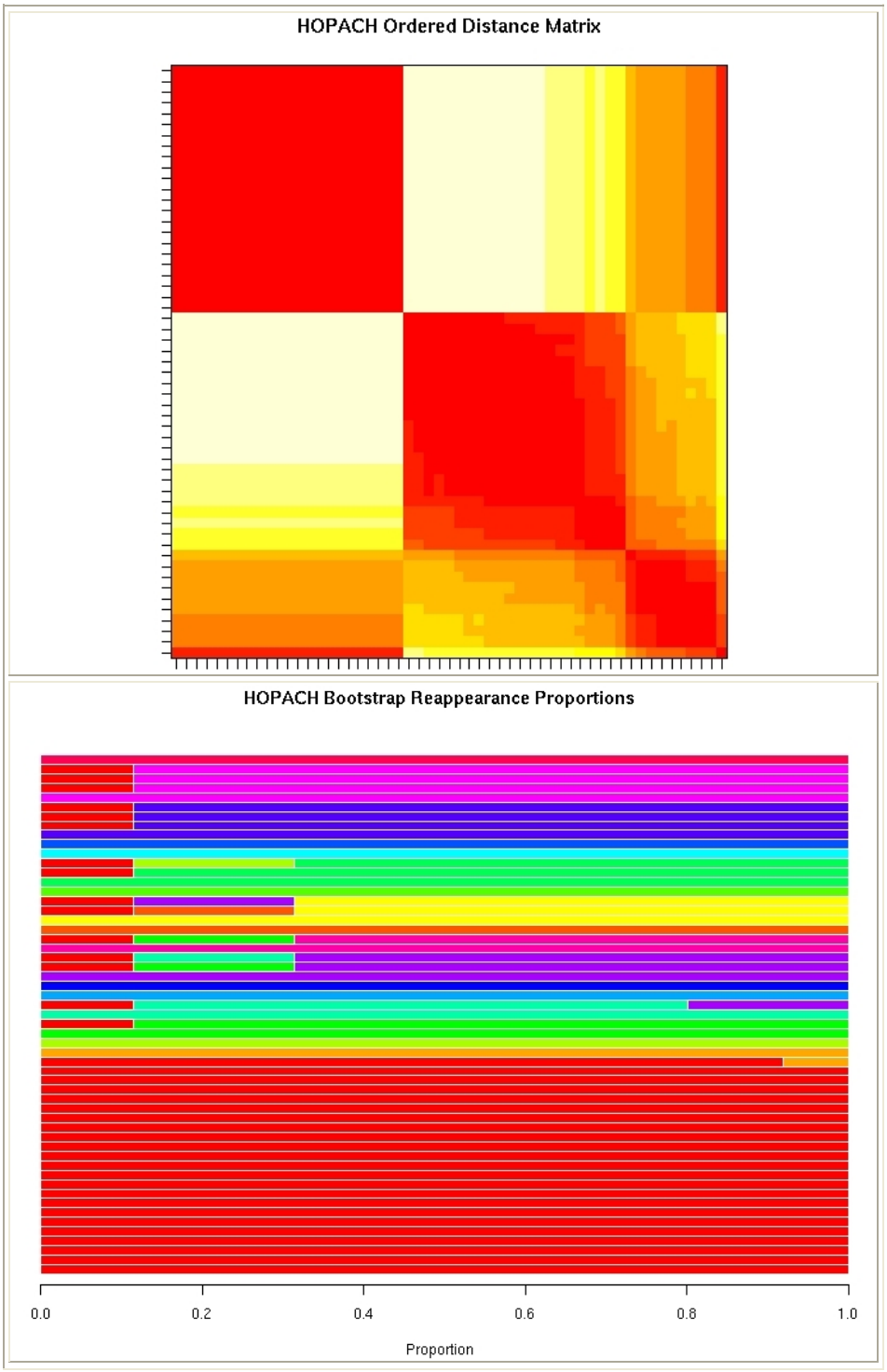


Figure 2.18: Ordered distance matrix and bootstrap reappearance proportions of genes with significant variance between the strains, but not within the strains in the skin treatment group.

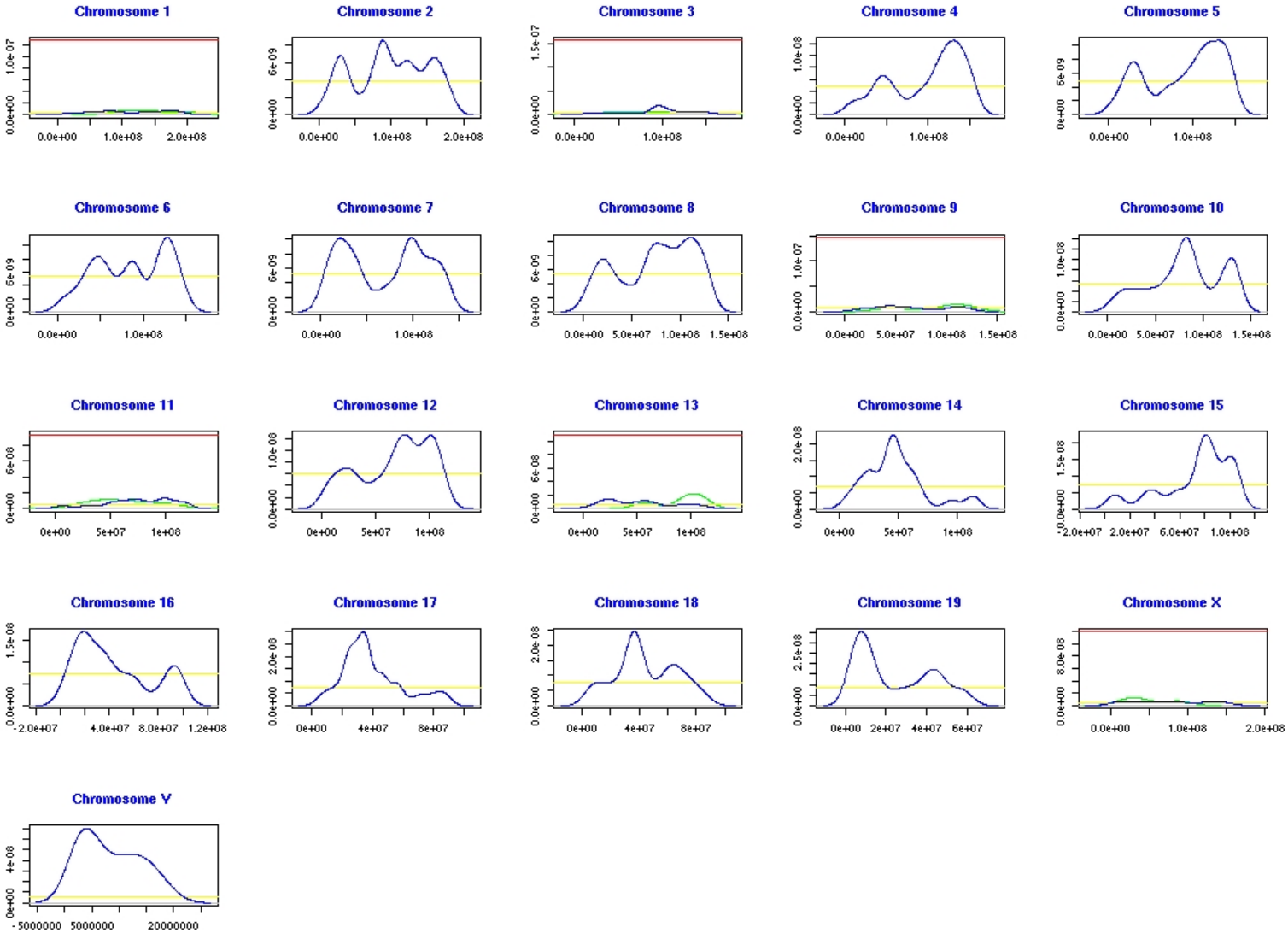


Figure 2.19: Gene density plots of genes with significant variance between the strains, but not within in the skin treatment group (within strain variance < 20,000 and between strain variance > 100,000).

6. Genes with significant variance between the strains, but not within the strain in the spleen treatment group (within strain variance < 20,000 and between strain variance > 100,000). There were 34 genes in this set, 3 of which didn't have a Gene ID. This reduced the set to 31 genes, all of which had an assigned chromosomal location. The gene density plots of each chromosome are in *Figure 2.20*. There were no gene-dense regions found on any of the chromosomes from this set. Since there were such a small number of genes in the set, clustering was not performed. 5 of these genes did not have an assigned Gene Ontology annotation. There was no Gene Ontology terms found to be enriched in this set.

7. Genes with a significant variance ratio in the treatment group, but not in the control group in the skin tissue, in 5 of the 6 strains. There were 865 genes in this set, 151 of which didn't have a Gene ID. This reduced the set to 714 genes, 96 of which did not have a chromosomal location. The gene density plots of each chromosome are in *Figure 2.22*. There were no gene-dense regions found on any of the chromosomes from this set.

Since this set was so large, clustering was performed using the HOPACH algorithm. There were three variables used for clustering: the average variance ratio for the strains, the maximum variance ratio and the range of the variance ratios. From this data, the HOPACH algorithm identified 90 main clusters in the 6^{th} level of the tree, with a median split silhouette of 0.345169. *Figure 2.21* shows the ordered distance matrix of the genes in this set. Many of the genes have a positive correlation (very little dissimilarity), although the inner branches don't seem to have any correlation with the outer branches of the HOPACH tree. *Figure 2.21* also shows the barplot of the bootstrap reappearance proportions. Interestingly, almost all of the genes have 10% of their reappearance proportion in the same cluster. Most genes belong to 3-4 clusters in this set.

Since there were so many clusters derived by HOPACH, clusters with less than 10 genes in them were combined into clusters of 20 or more genes, as described in the Methods section. This yielded 81 clusters that had at least 10 genes each in them. From this set, there were 159 genes that did not have Gene Ontology annotations. Each cluster was analyzed for enriched terms, as described in the Methods section, and four terms were found to be enriched: male sex differentiation (biological process), integrin-mediated signaling pathway (biological process), external side of plasma membrane (cellular components) and cell surface (cellular components).

8. Genes with a significant variance ratio in the treatment group, but not in the control group in the spleen tissue, in 5 of 6 strains. There were 146 genes in this set, 12 of which did not have a Gene ID. This reduced the set to 134 genes, 12 of which did not have a chromosomal location. The gene density plots of each chromosome are in *Figure 2.24*. There were no gene-dense regions found on any of the chromosomes from this set. Since this set was so large, clustering was done using the HOPACH algorithm. There were three variables used for clustering: the average variance ratio for the strains, the maximum variance ratio and the range of the variance ratios. From this data, the HOPACH algorithm identified 52 main clusters in the 3^{th} level of the tree, with a median split silhouette of 0.2230841. *Figure 2.23* shows the ordered distance matrix of the genes

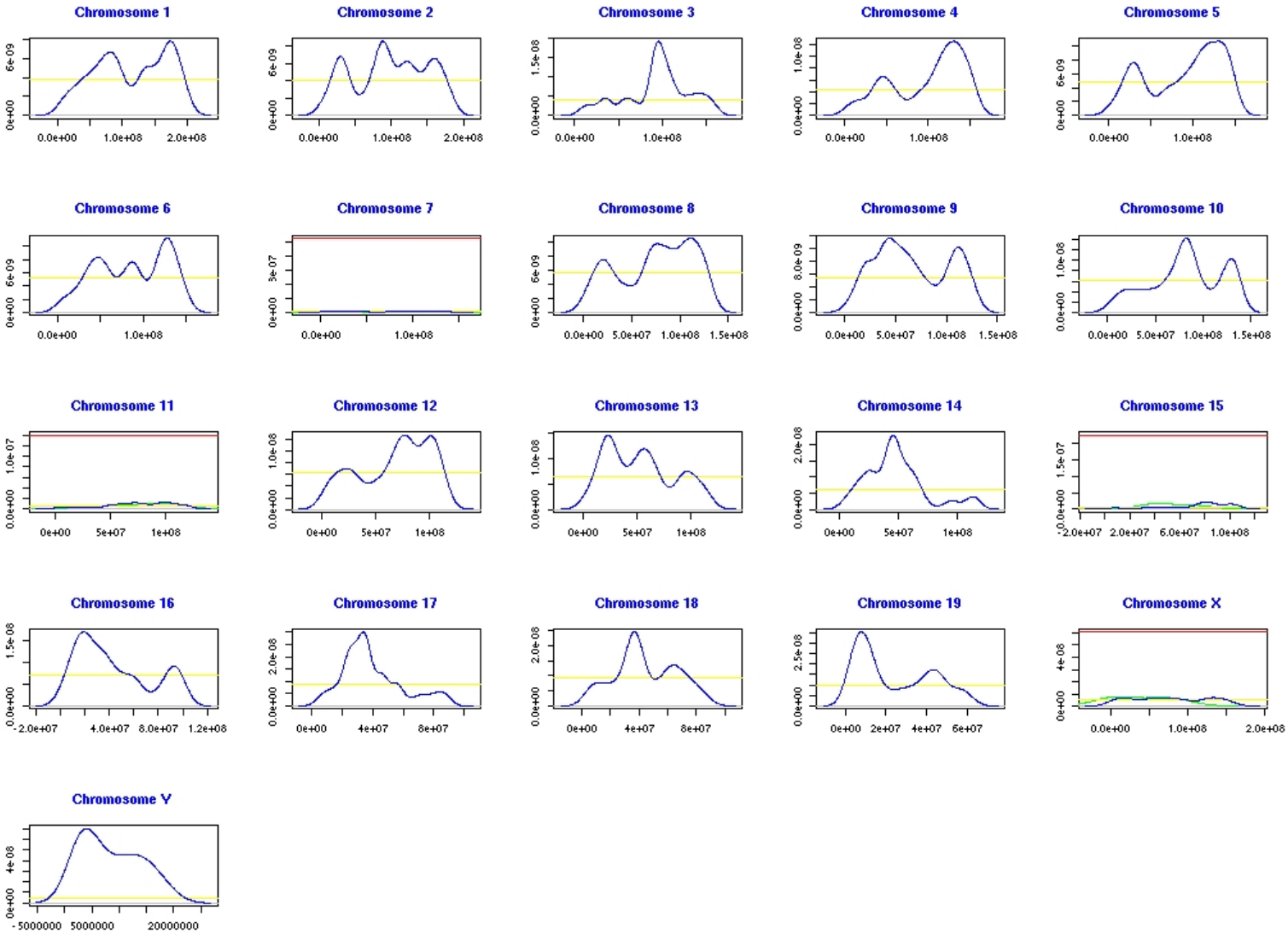


Figure 2.20: Gene density plots of genes with significant variance between the strains, but not within in the spleen treatment group (within strain variance < 20,000 and between strain variance > 100,000).

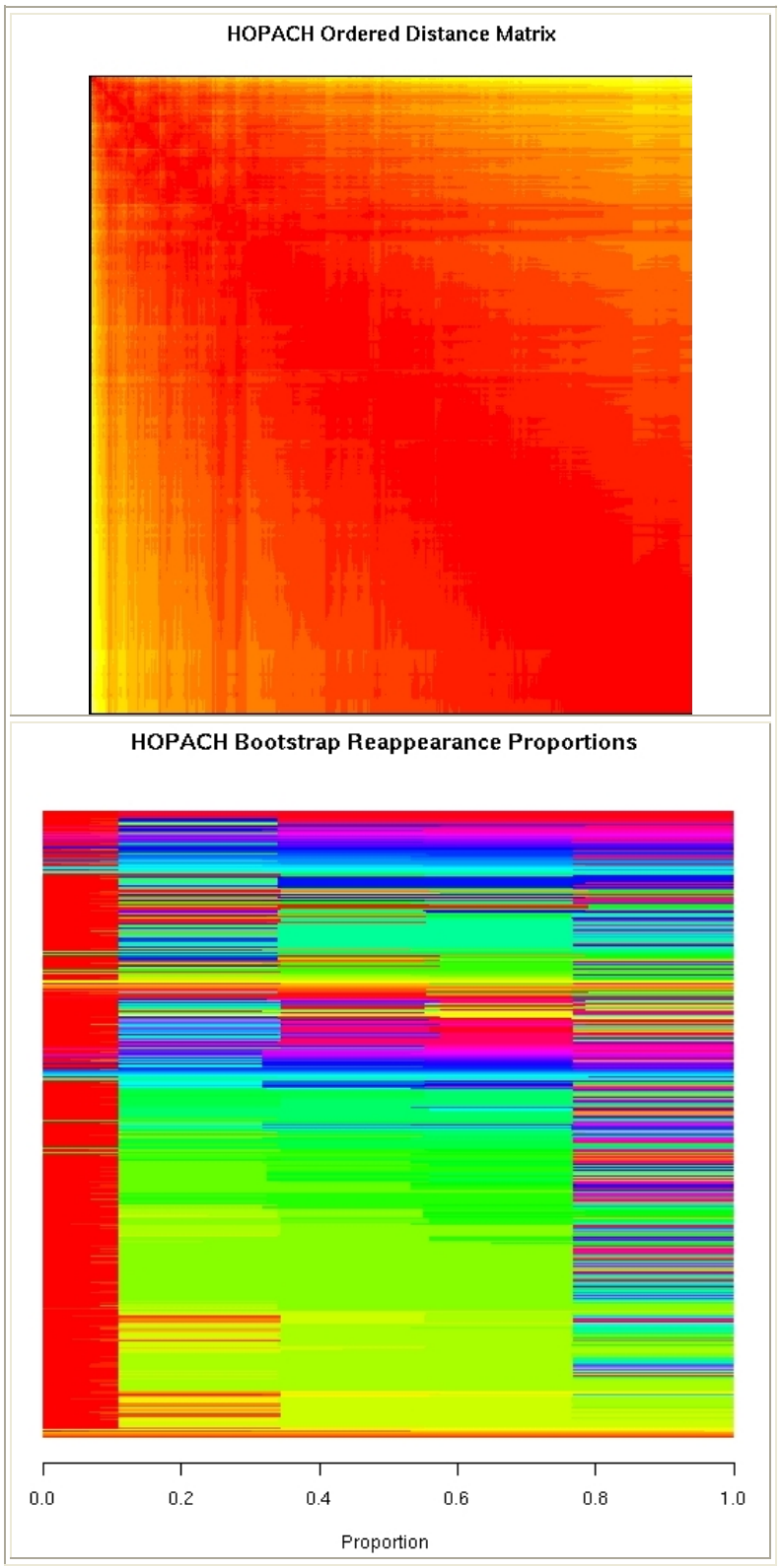


Figure 2.21: Ordered distance matrix and bootstrap reappearance proportions of genes with a significant variance ratio in the treatment group, but not in the control group in the skin tissue, in 5 of 6 strains.

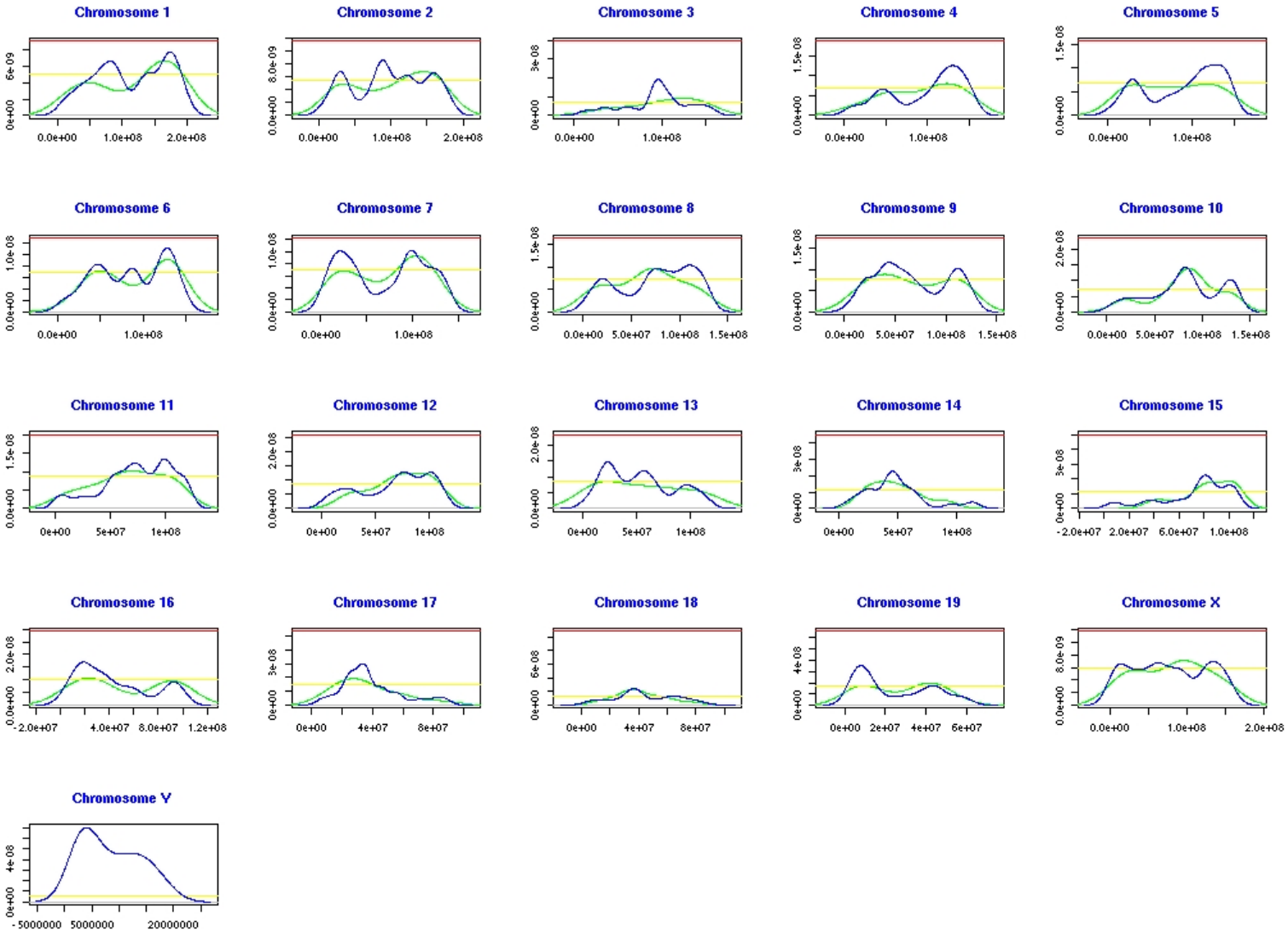


Figure 2.22: Gene density plots of genes with a significant variance ratio in the treatment group, but not in the control group in the skin tissue, in 5 of 6 strains.

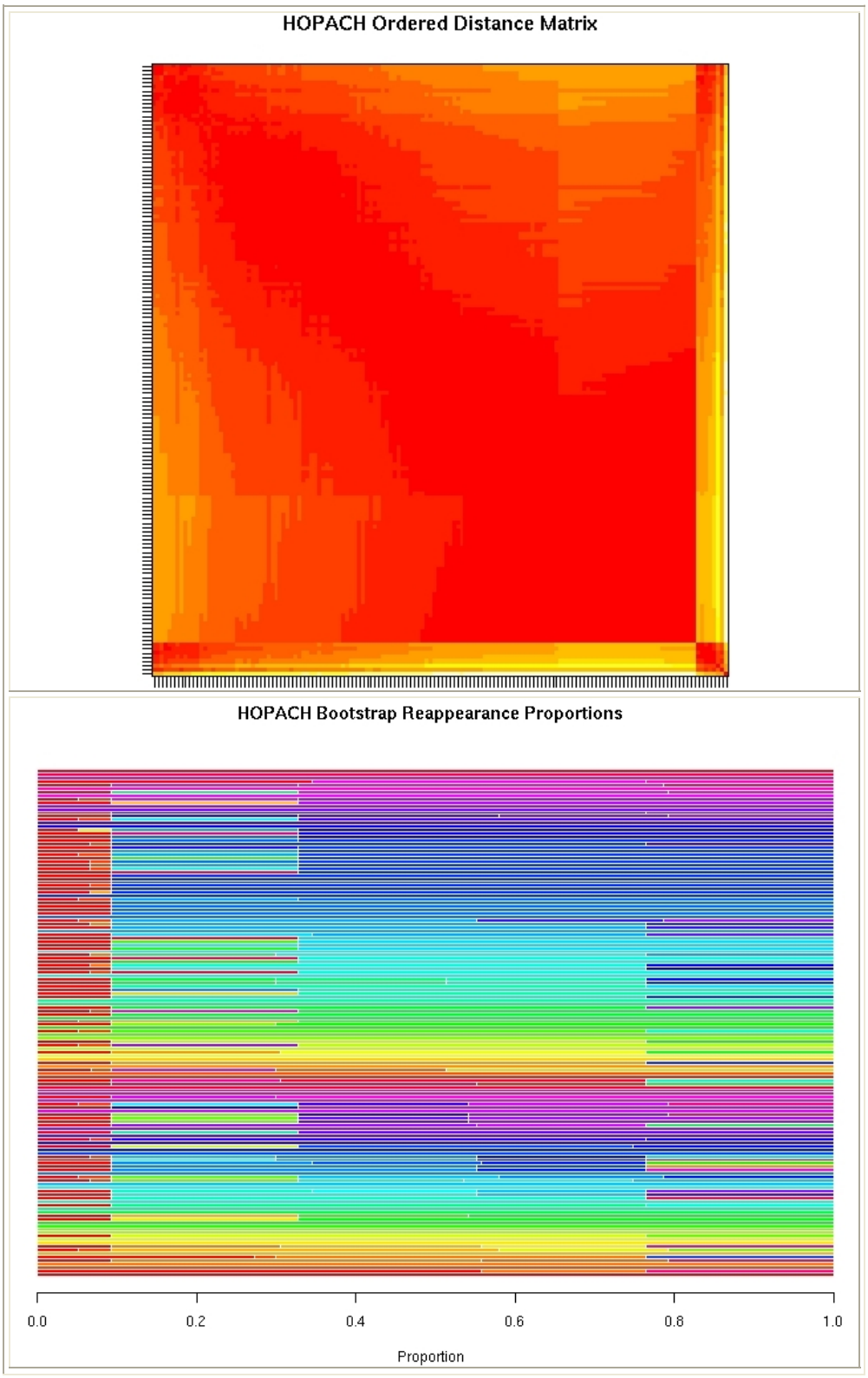


Figure 2.23: Ordered distance matrix and bootstrap reappearance proportions of genes with a significant variance ratio in the treatment group, but not in the control group in the spleen tissue, and common between 5 of 6 strains.

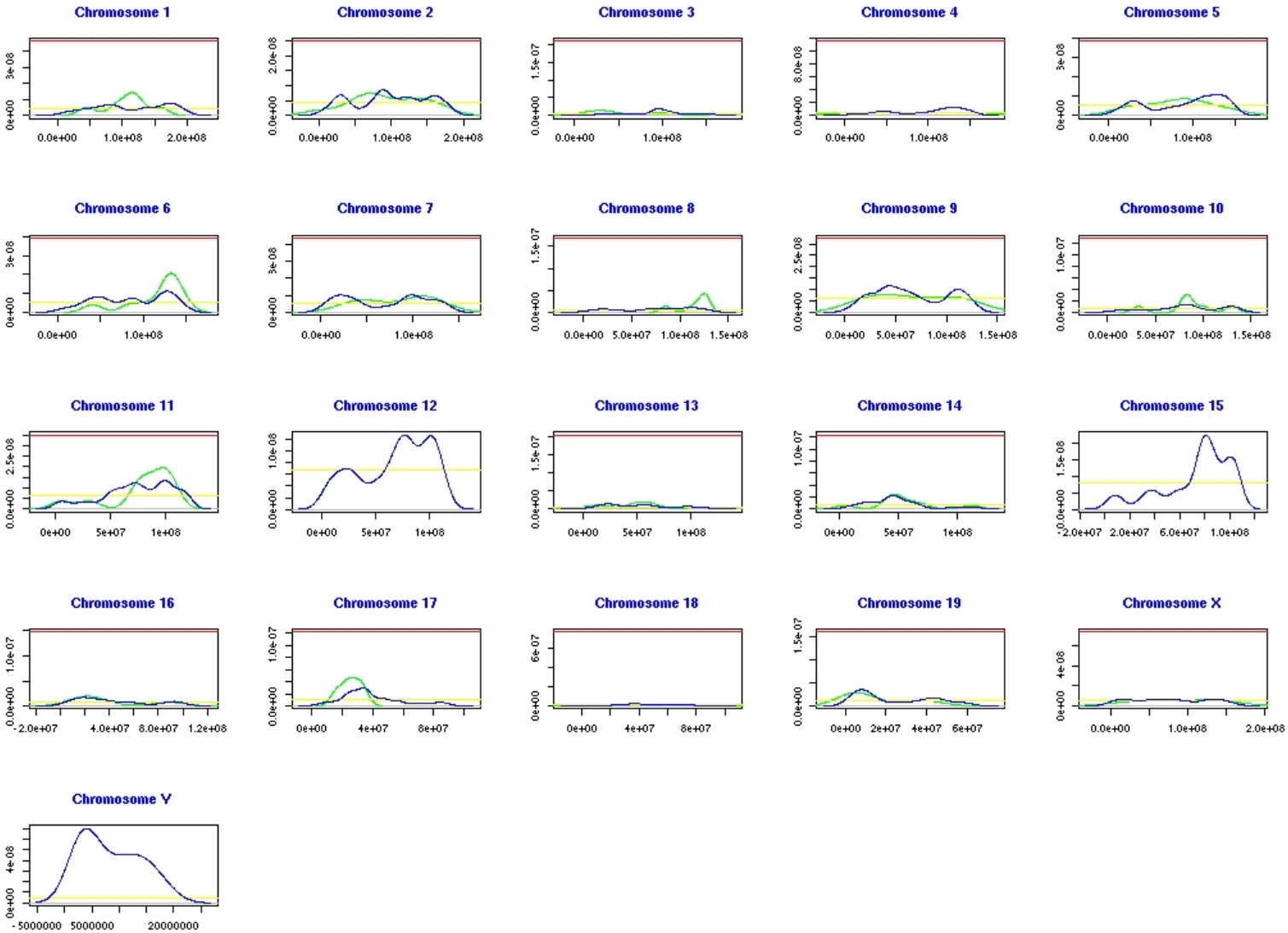


Figure 2.24: Gene density plots of genes with a significant variance ratio in the treatment group, but not in the control group in the spleen tissue, in 5 of 6 strains.

in this set. Almost all of the genes have a positive correlation (very little dissimilarity), although there are some in the outer branches of the HOPACH tree that have no correlation. *Figure 2.23* also shows the barplot of the bootstrap reappearance proportions. Most of the genes in this set below to three different clusters, although genes in the first 5 HOPACH clusters belong to only one cluster.

Since there were so many clusters derived by HOPACH, clusters with less than 10 genes in them were combined into clusters of 20 or more genes, as described in the Methods section. This yielded 15 clusters that had at least 10 genes each in them. From this set, there were 9 genes that did not have Gene Ontology annotations. Each cluster was analyzed for enriched terms, as described in the Methods section, but no terms were found to be enriched in this set.

9. Genes with a significant variance ratio in the control and treatment groups, and differential expression between these groups in the skin tissue. There were 689 genes in the set, 51 of which did not have a Gene ID. This reduced the set to 638 genes, 288 of which did not have an assigned chromosomal location. The gene density plots of each chromosome are in *Figure 2.25*. There were two regions found to be gene-dense on the genome: one on chromosome 3 and the other on chromosome 11.

The peak of the density curve on chromosome 3 spans 2.5 megabases, in which there are 9 genes from the 12 genes in the set on the chromosome. The genes under the curve are the following: RIKEN cDNA 6330415M09 gene, farnesyl diphosphate synthetase, RIKEN cDNA 4932431F02 gene, RIKEN cDNA 4932431F02 gene, RIKEN cDNA 1110058A15 gene, mitochondrial ribosomal protein L9, RIKEN cDNA 2310007A19 gene, RIKEN cDNA 2610022K04 gene and CDC42 small effector 1. These genes had start sites of 88863530, 91299735, 92413789, 92413789, 95637164, 97141147, 97162792, 97835296 and 97929318 respectively. Four of these 9 genes were not annotated with Gene Ontology categories, and no categories were found to be enriched.

The peak of the density curve on chromosome 11 spans 5 megabases, in which there are 14 genes from the 22 genes on the chromosome. The genes under the curve are the following: peripheral myelin protein, RIKEN cDNA 4632403M07 gene, RIKEN cDNA 1700095G12 gene, solute carrier family 25 (mitochondrial carrier oxoglutarate carrier) member 11, MIS12 homolog (yeast), RIKEN cDNA 1200014J11 gene, carbohydrate kinase-like, RIKEN cDNA 0610007A15 gene, DNA segment Chr 11, ERATO Doi 530 expressed, RIKEN cDNA 5730455P16 gene, acetyl-Coenzyme A carboxylase, RIKEN cDNA 1110001A07 gene, RIKEN cDNA 4732452J19 gene, and splicing factor arginine/serine-rich 1 (ASF/SF2). These genes had start sites of 62702778, 69374882, 69392138, 70215466, 70590877, 72619145, 72770747, 77750031, 79564371, 79932720, 83786523, 86669632, 87260268 and 87608150 respectively. Six of these 14 genes were not annotated with a Gene Ontology category, and zero categories were enriched in the remaining 8 genes.

Since this set was so large, clustering was performed using the HOPACH algorithm. There were three variables used for clustering: the t-test value, the variance ratio from the control group and the variance ratio from the treatment group. From this data, the HOPACH algorithm identified 81 main clusters in the 6^{th} level of the tree, with

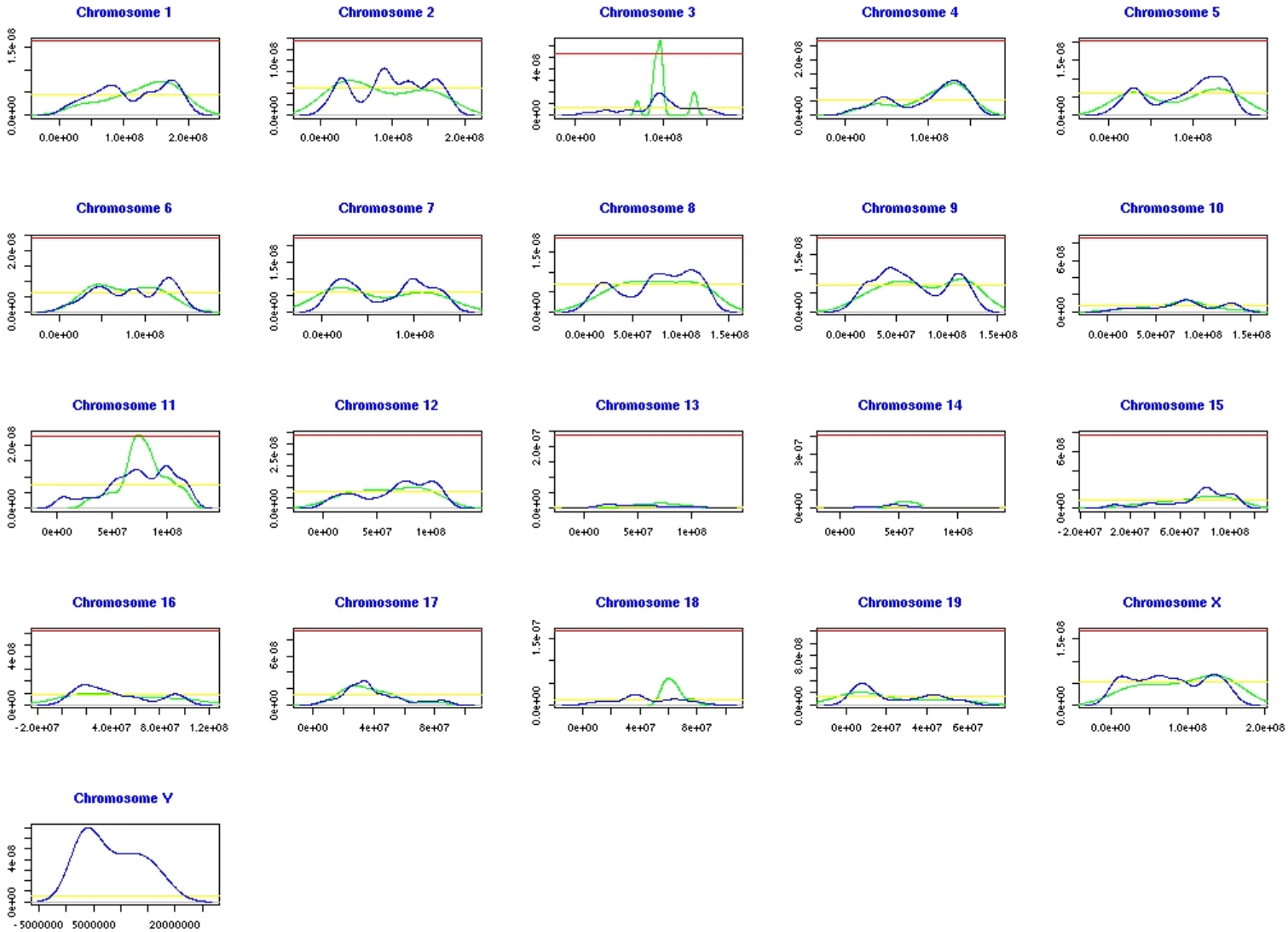


Figure 2.25: Gene density plots of genes with a significant variance ratio in the control and treatment groups, and was differentially expressed between these groups in the skin tissue.

a median split silhouette of 0.2742538. *Figure 2.26* shows the ordered distance matrix of the genes in this set. Some of the genes in the middle branches of the HOPACH tree positively correlate with each other, but the outer and inner-most branches of the tree don't correlate with themselves or each other. *Figure 2.26* also shows the barplot of the bootstrap reappearance proportions. Almost all of the genes belong to the first HOPACH cluster. Most of the genes are in at least 5 different clusters, and only a few belong exclusively to one cluster.

Since there were so many clusters derived by HOPACH, clusters with less than 10 genes in them were combined into clusters of 20 or more genes, as described in the Methods section. This yielded 68 clusters that had at least 10 genes each in them. From this set, there were no genes that did not have Gene Ontology annotations. Each cluster was analyzed for enriched terms, as described in the Methods section, but no terms were found to be enriched in this set.

10. Genes that were significantly differentially expressed between the control and treatment groups, in all six strains in the skin tissue. There were 1,563 genes in this set, 233 of which did not have a Gene ID. This reduced the set to 1,330 genes, 372 of which did not have an assigned chromosomal location. The gene density plots of each chromosome are in *Figure 2.28*. There were no gene-dense regions found on any of the chromosomes from this set.

Since this set was so large, clustering was performed using the HOPACH algorithm. There were three variables used for clustering: the average t-value for the strains, the maximum t-value and the range of the t-values. From this data, the HOPACH algorithm identified 81 main clusters in the 6^{th} level of the tree, with a median split silhouette of 0.2742538. *Figure 2.27* shows the ordered distance matrix of the genes in this set. A majority of the clusters positively correlate with each other, although the outer branches of the HOPACH tree negatively correlate, while the branches just inside of those don't correlate at all with each other. *Figure 2.27* also shows the barplot of the bootstrap reappearance proportions. Almost all of the genes belong to the first HOPACH cluster. Most of the genes are in at least 5 different clusters, and none of them belong exclusively to one cluster. There are a group of genes that have only a small reappearance proportion in the first cluster, then the remainder in another common cluster.

Since there were so many clusters derived by HOPACH, clusters with less than 10 genes in them were combined into clusters of 20 or more genes, as described in the Methods section. This yielded 117 clusters that had at least 10 genes each in them. From this set, there were 471 genes that did not have Gene Ontology annotations. Each cluster was analyzed for enriched terms, as described in the Methods section, and only one term was found to be significantly enriched: molecular function unknown (molecular function).

11. Genes that were significantly differentially expressed between the control and treatment groups, in all six of the strains in the spleen tissue. There were 655 genes in this set, 98 of which did not have a Gene ID. This reduced the set to 557 genes, 49 of which did not have an assigned chromosomal location. The gene density plots of each

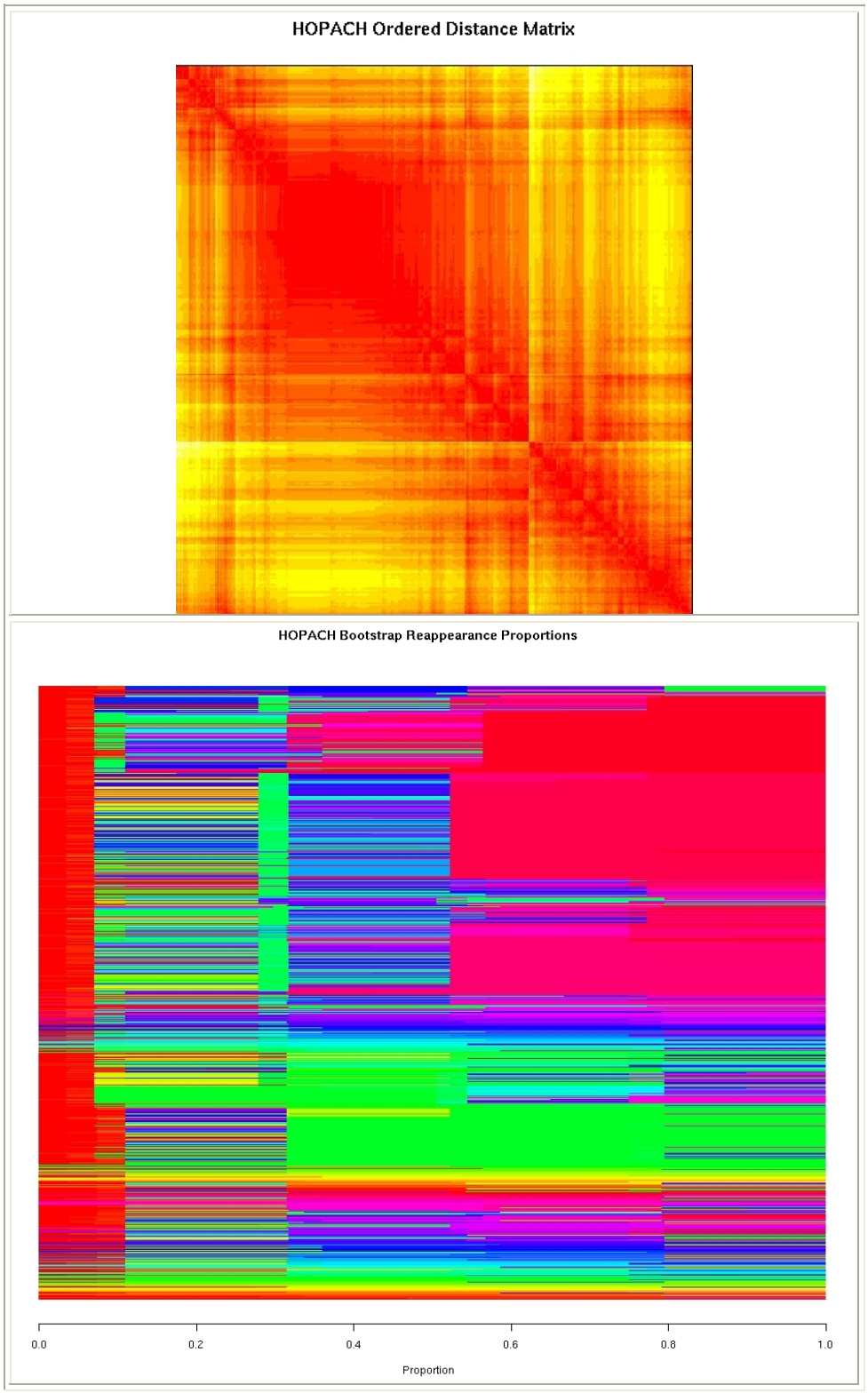


Figure 2.26: Ordered distance matrix and bootstrap reappearance proportions of genes with a significant variance ratio in the control and treatment groups, and was differentially expressed between these groups in the skin tissue.

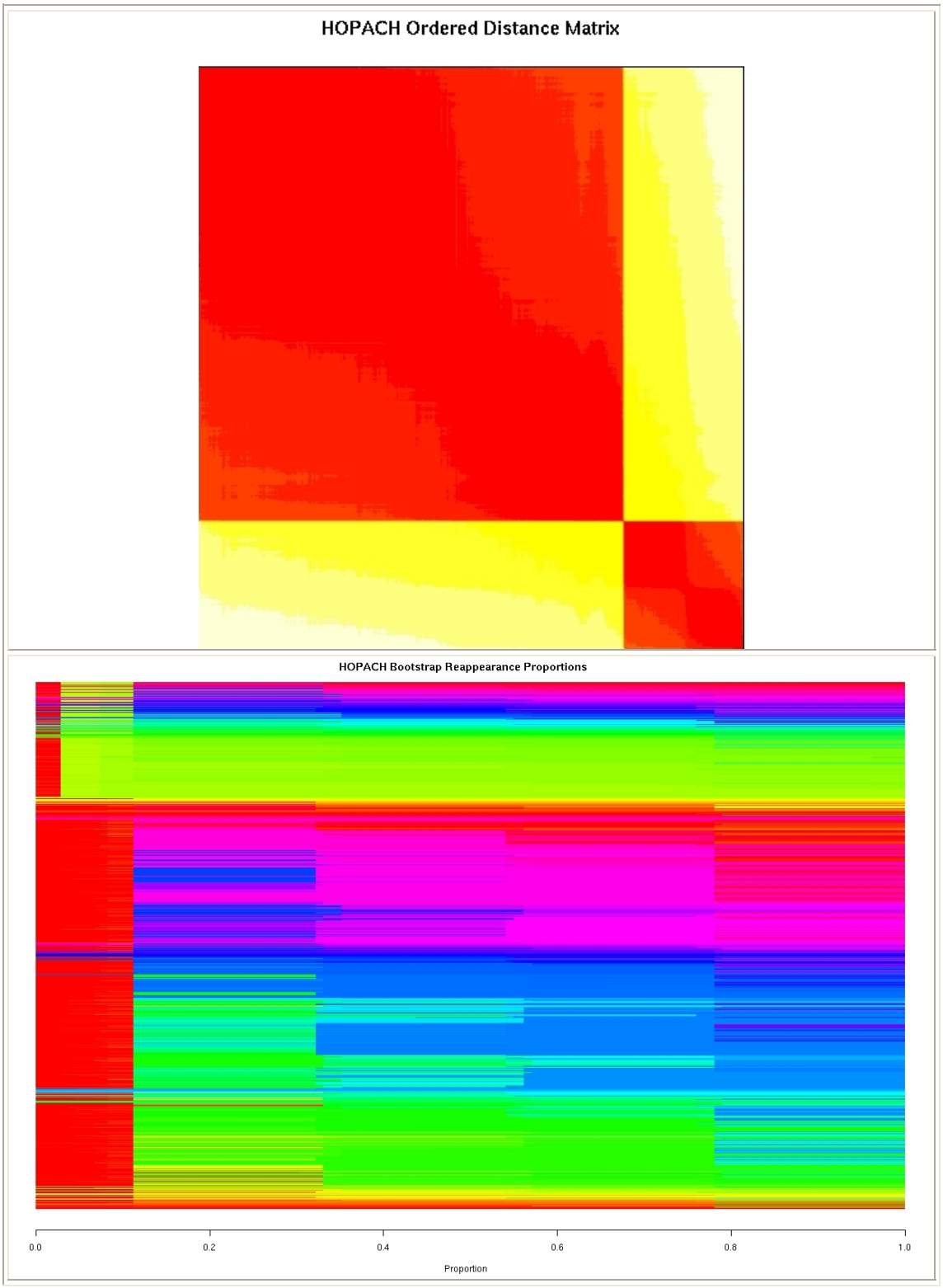


Figure 2.27: Ordered distance matrix and bootstrap reappearance proportions of genes that were significantly differentially expressed between the control and treatment groups, in all of the strains in the skin tissue.

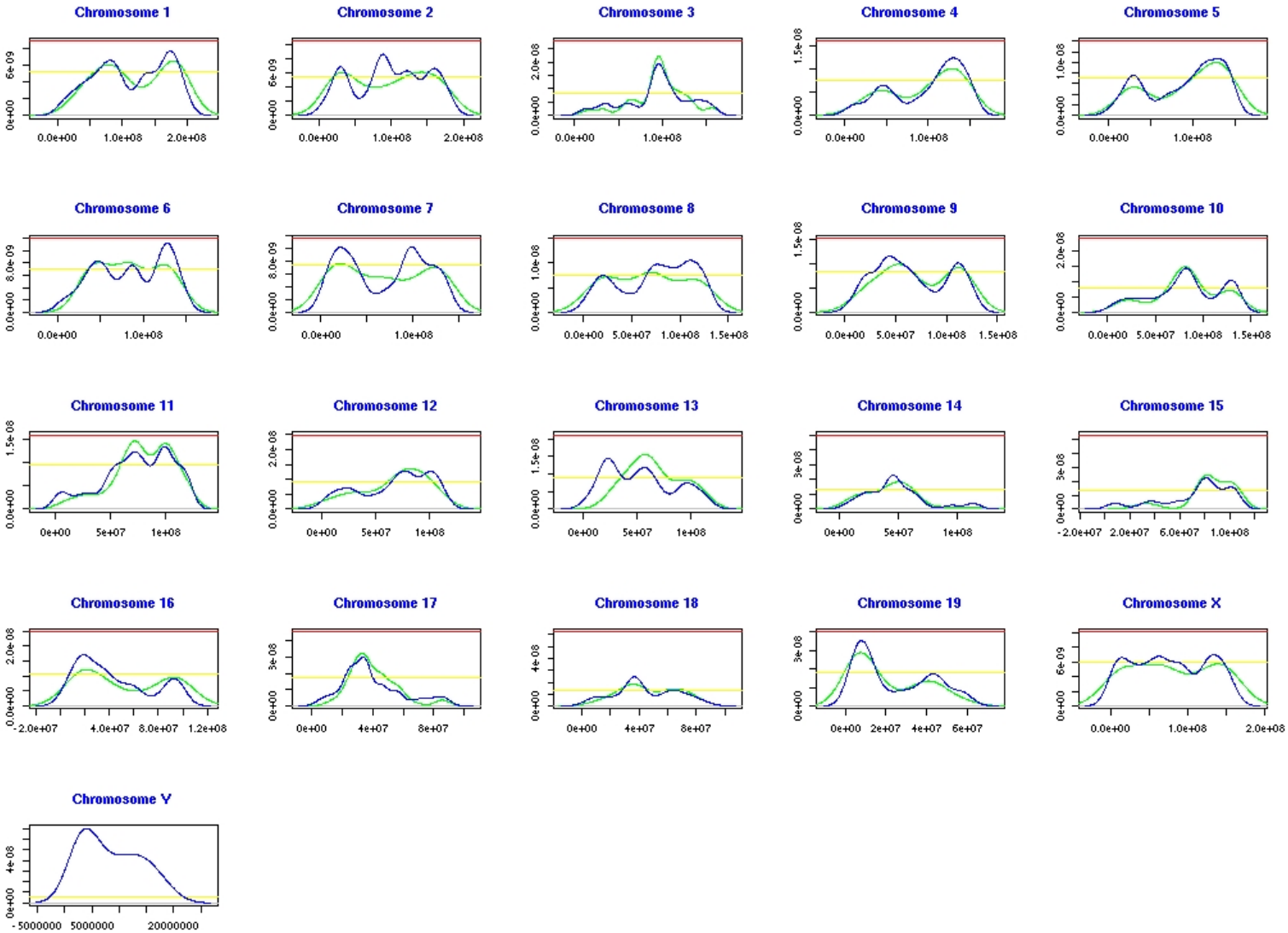


Figure 2.28: Gene density plots of genes significantly differentially expressed between the control and treatment groups, in all of the strains in the skin tissue.

chromosome are in *Figure 2.30*. There were no gene-dense regions found on any of the chromosomes from this set.

Since this set was so large, clustering was performed using the HOPACH algorithm. There were three variables used for clustering: the average t-value for the strains, the maximum t-value and the range of the t-values. From this data, the HOPACH algorithm identified 102 main clusters in the 6th level of the tree, with a median split silhouette of 0.2677498. *Figure 2.29* shows the ordered distance matrix of the genes in this set. A majority of the clusters positively correlate with each other, although the outer branches of the HOPACH tree negatively correlate, while the branches just inside of those don't correlate at all with each other. *Figure 2.29* also shows the barplot of the bootstrap reappearance proportions. Many of the genes belong to the first HOPACH cluster. Most of the genes are in 2 or 3 different clusters, and none belong exclusively to one cluster.

Since there were so many clusters derived by HOPACH, clusters with less than 10 genes in them were combined into clusters of 20 or more genes, as described in the Methods section. This yielded 53 clusters that had at least 10 genes each in them. From this set, there were 75 genes that did not have Gene Ontology annotations. Each cluster was analyzed for enriched terms, as described in the Methods section, but no terms were found to be enriched in this set.

12. Genes that had a significant variance ratio in one and only one strain in the skin control group. There were 3,023 genes in this set, 295 of which did not have a Gene ID. This reduced the set to 2,728 genes, 1,056 of which did not have an assigned chromosomal location. The gene density plots of each chromosome are in *Figure 2.32*. There were no gene-dense regions found on any of the chromosomes from this set.

Since this set was extremely large, clustering was performed using the HOPACH algorithm. There was only one variable available for clustering with this set, which was the variance ratio for each significant gene in the strain. From this data, the HOPACH algorithm identified 2 main clusters in the 1st level of the tree, with a median split silhouette of -0.3825825. *Figure 2.31* shows the ordered distance matrix of the genes in this set. There are no major trends of clusters not correlating with each other – it appears that almost all of the clusters have a positive correlation with one another. *Figure 2.31* also shows the barplot of the bootstrap reappearance proportions. From this plot, all of the genes belong entirely to the same cluster, so bootstrap had no effect on the data.

Since both of the clusters found by HOPACH had more than 10 genes in them, no new clusters needed to be created. There were 1,435 genes in this set that were not annotated with Gene Ontology terms, and no terms were found to be enriched in either cluster.

13. Genes that had a significant variance ratio in one and only one strain in the spleen control group. There were 3,019 genes in this set, 364 of which did not have a Gene ID. This reduced the set to 2,655 genes, 263 of which did not have an assigned chromosomal location. The gene density plots of each chromosome are in *Figure 2.33*. There were no gene-dense regions found on any of the chromosomes from this set.

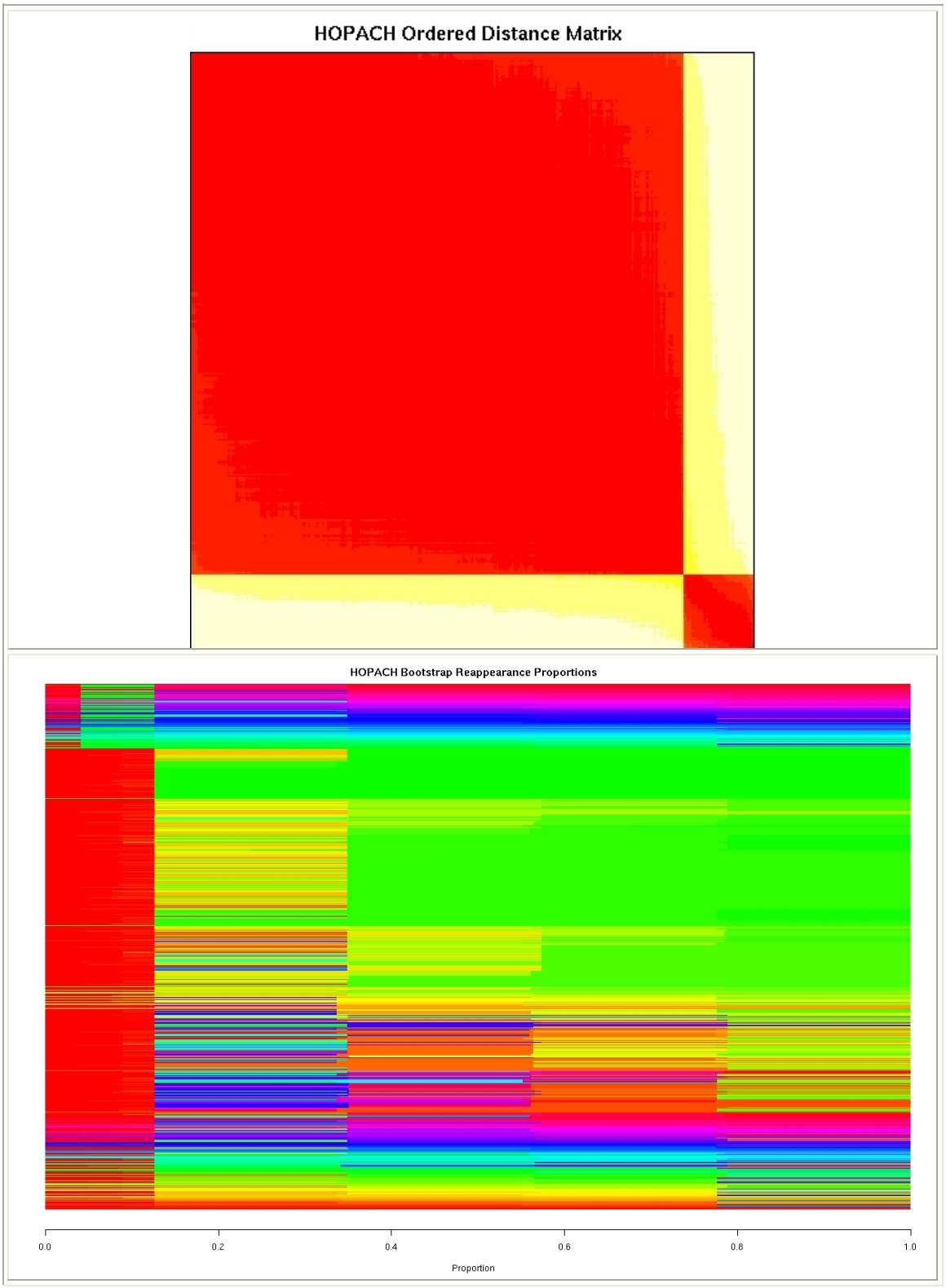


Figure 2.29: Ordered distance matrix and bootstrap reappearance proportions of genes that were significantly differentially expressed between the control and treatment groups, in all of the strains in the spleen tissue.

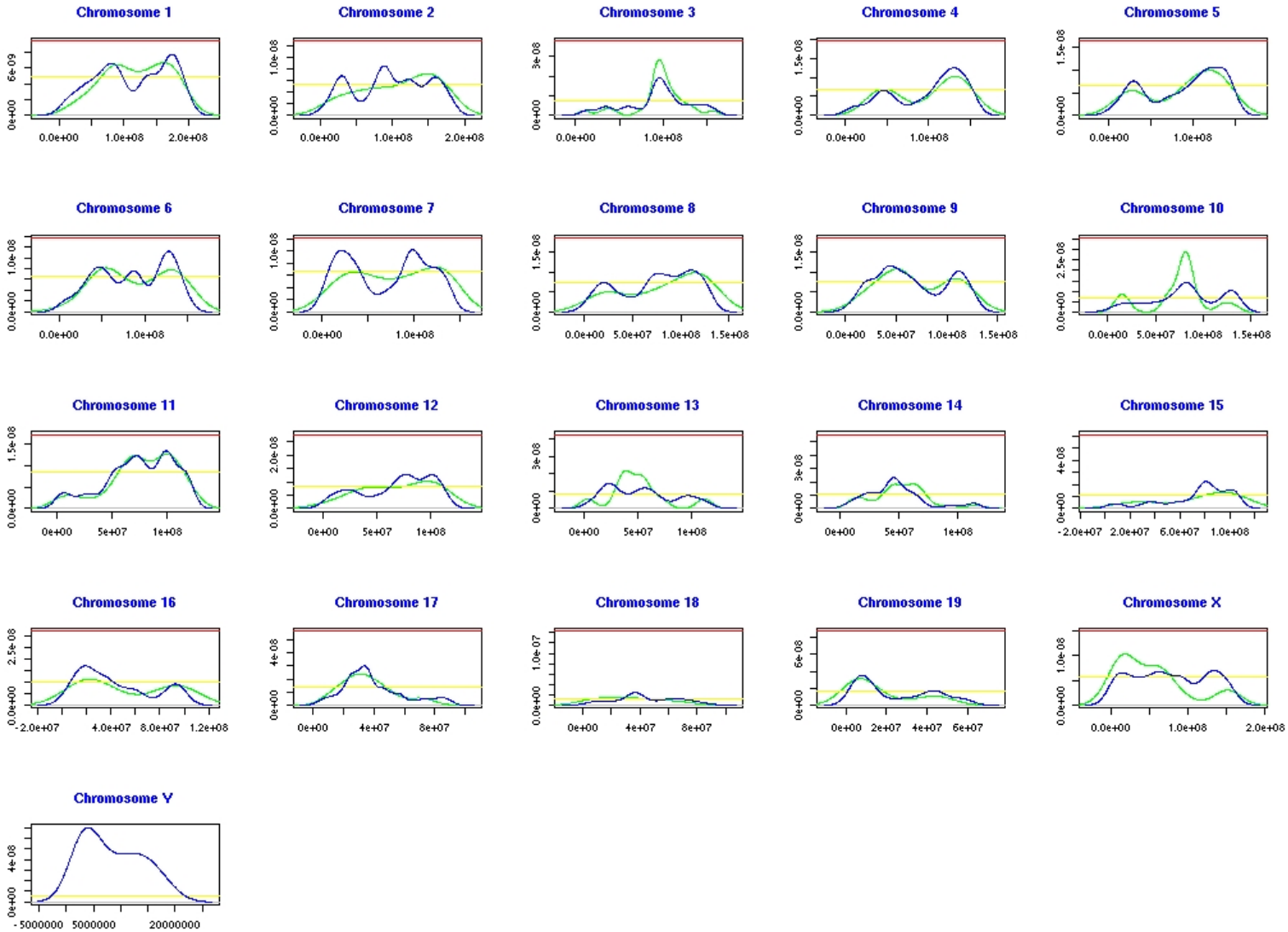


Figure 2.30: Gene density plots of genes significantly differentially expressed between the control and treatment groups, in all 6 strains in the spleen tissue.

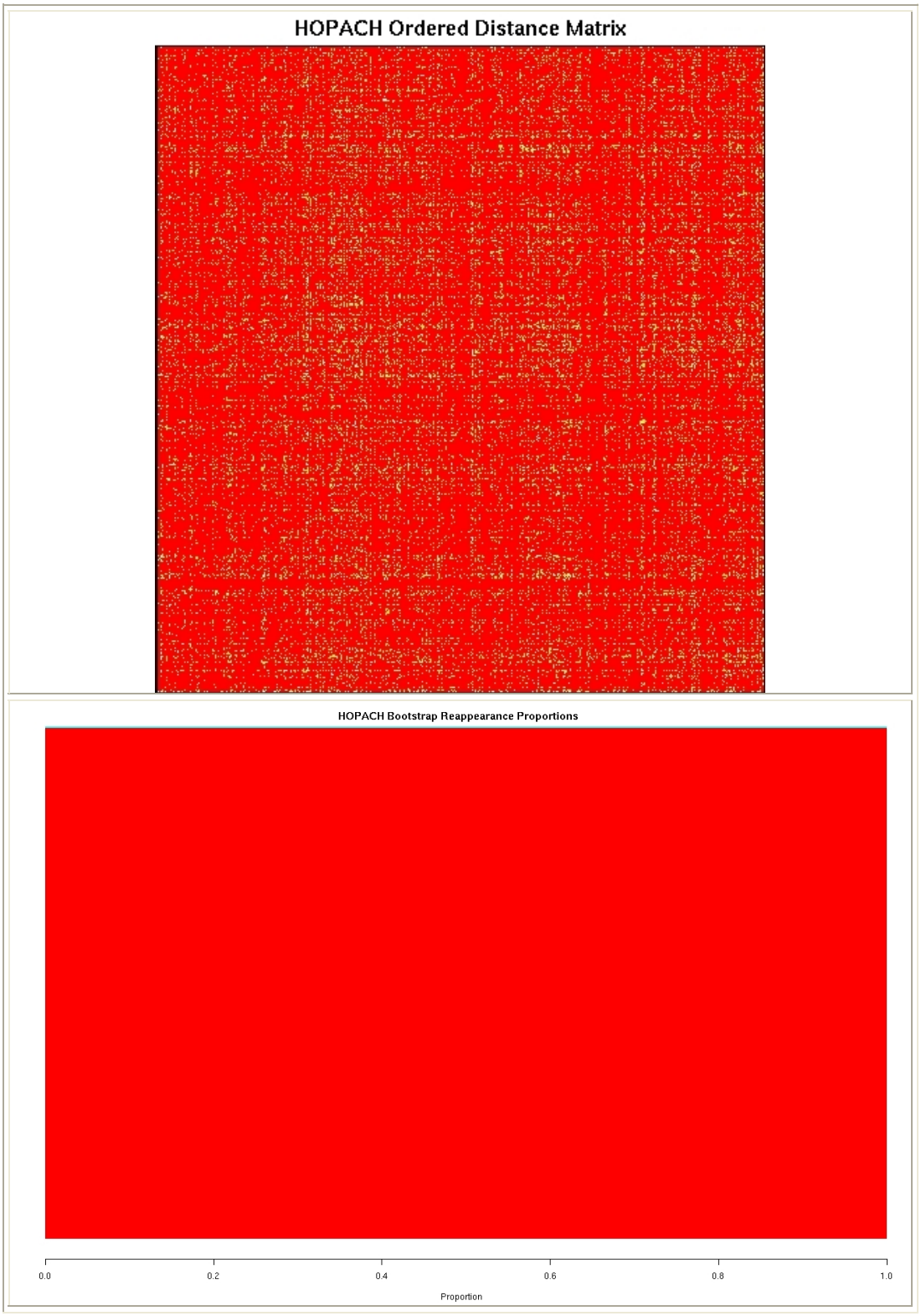


Figure 2.31: Ordered distance matrix and bootstrap reappearance proportions of genes that had a significant variance ratio in one and only one strain in the skin control group.

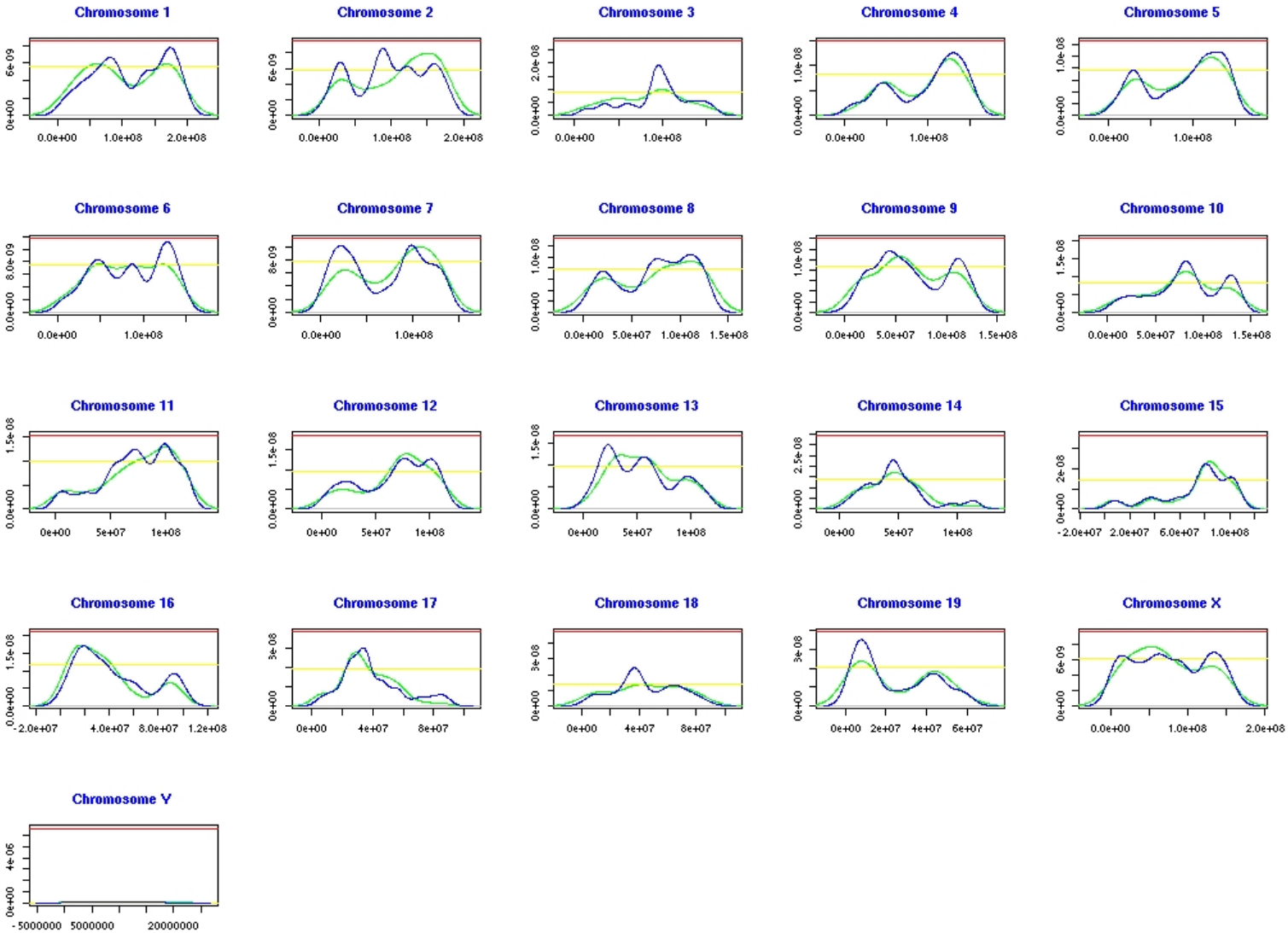


Figure 2.32: Gene density plots of genes that have a significant variance ratio in one and only one strain in the skin control group.

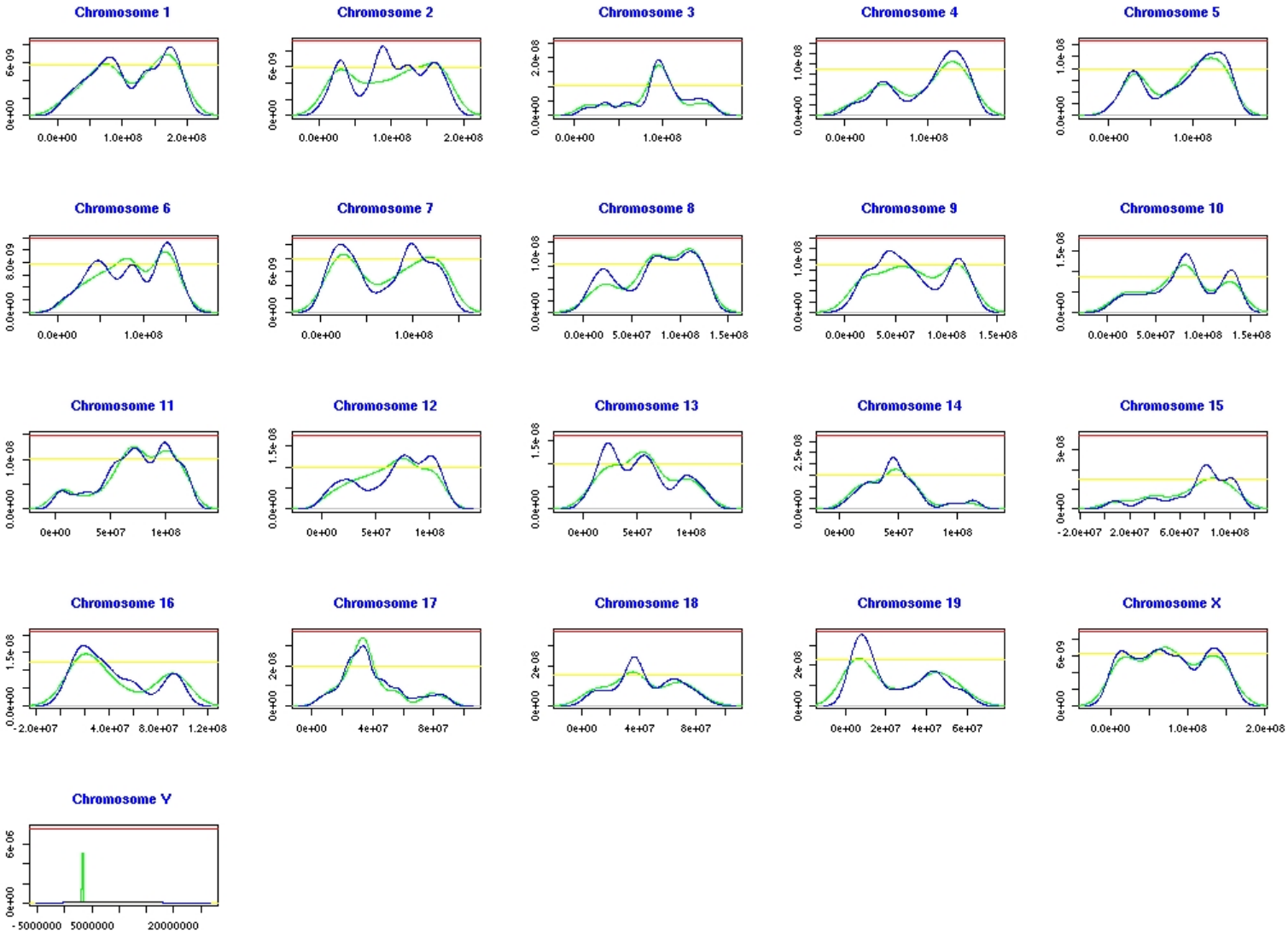


Figure 2.33: Gene density plots of genes that have a significant variance ratio in one and only one strain in the spleen control group.

Since this set was extremely large, clustering was done using the HOPACH algorithm. There was only one variable available for clustering with this set, which was the variance ratio for each significant gene in the strain. From this data, the HOPACH algorithm identified 2 main clusters in the 1st level of the tree, with a median split silhouette of -0.3333333. *Figure 2.34* shows the ordered distance matrix of the genes in this set. There are no major trends of clusters not correlating with each other – it appears that almost all of the clusters have a positive correlation with one another. *Figure 2.34* also shows the barplot of the bootstrap reappearance proportions. From this plot, all of the genes belong entirely to the same cluster, so bootstrap had no effect on the data.

Since both of the clusters found by HOPACH had more than 10 genes in them, no new clusters needed to be created. There were 441 genes in this set that were not annotated with Gene Ontology terms, and no terms were found to be enriched in either cluster.

14. Genes that had a significant variance ratio in one and only one strain in the skin treatment group. There were 4,663 genes in this set, 624 of which did not have a Gene ID. This reduced the set to 4,039 genes, 1,133 of which did not have an assigned chromosomal location. The gene density plots of each chromosome are in *Figure 2.36*. There were no gene-dense regions found on any of the chromosomes from this set.

Since this set was extremely large, clustering was done using the HOPACH algorithm. There was only one variable available for clustering with this set, which was the variance ratio for each significant gene in the strain. From this data, the HOPACH algorithm identified 2 main clusters in the 1st level of the tree, with a median split silhouette of -0.3928058. *Figure 2.35* shows the ordered distance matrix of the genes in this set. There are no major trends of clusters not correlating with each other – it appears that almost all of the clusters have a positive correlation with one another. *Figure 2.35* also shows the barplot of the bootstrap reappearance proportions. From this plot, all of the genes belong entirely to the same cluster, so bootstrap had no effect on the data.

Since both of the clusters found by HOPACH had more than 10 genes in them, no new clusters needed to be created. There were 1,638 genes in this set that were not annotated with Gene Ontology terms, and no terms were found to be enriched in either cluster.

15. Genes that had a significant variance ratio in one and only one strain in the spleen treatment group. There were 2,716 genes in this set, 341 of which did not have a Gene ID. This reduced the set to 2,375 genes, 234 of which did not have an assigned chromosomal location. The gene density plots of each chromosome are in *Figure 2.38*. There were no gene-dense regions found on any of the chromosomes from this set. Since this set was extremely large, clustering was performed using the HOPACH algorithm. There was only one variable available for clustering with this set, which was the variance ratio for each significant gene in the strain. From this data, the HOPACH algorithm identified 2 main clusters in the 1st level of the tree, with a median split silhouette of -0.16666667. *Figure 2.37* shows the ordered distance matrix of the genes in this set. There are no major trends of clusters not correlating with each other – it appears that almost all of the clusters have a positive correlation with one another. *Figure 2.37*

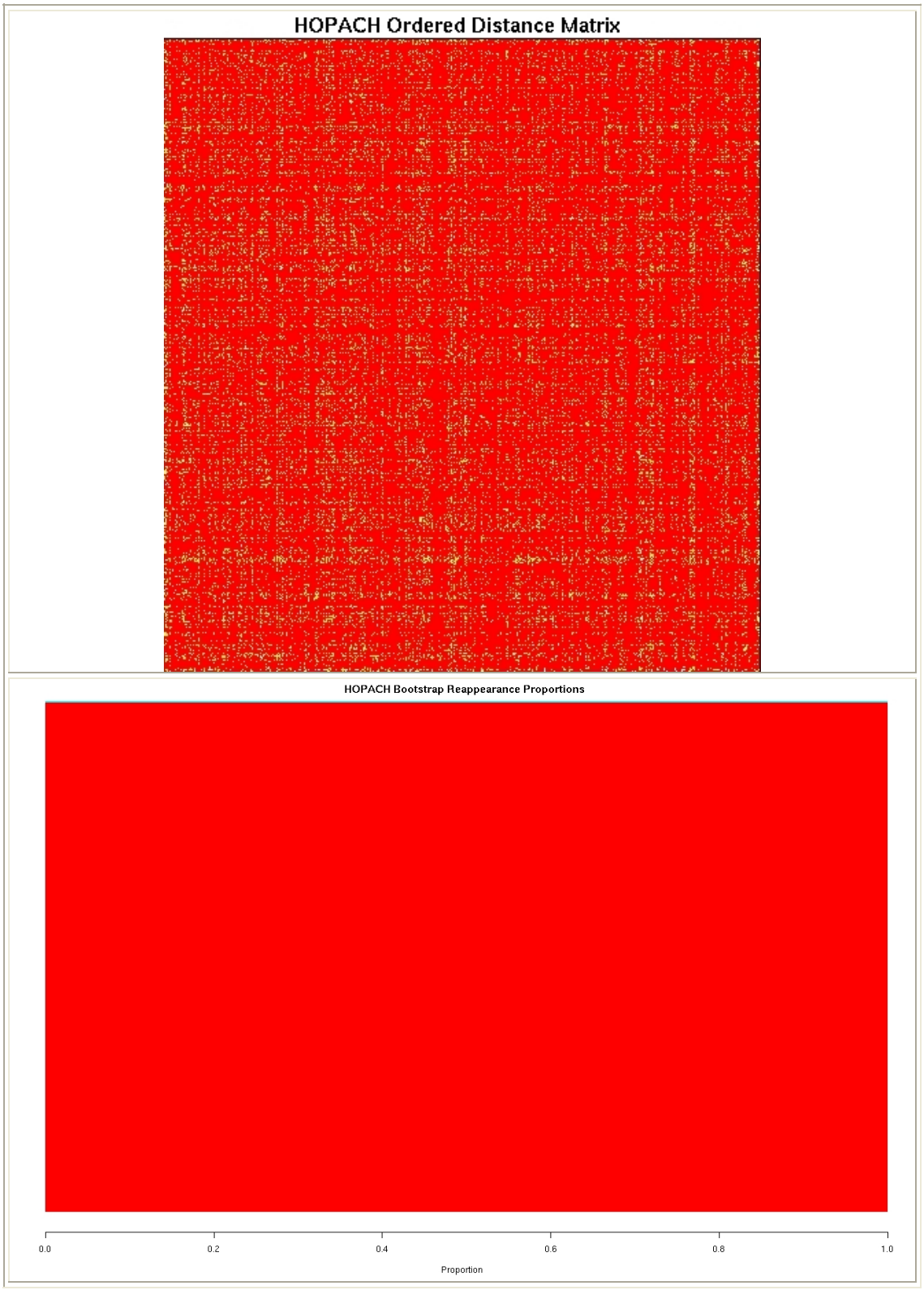


Figure 2.34: Ordered distance matrix and bootstrap reappearance proportions of genes that had a significant variance ratio in one and only one strain in the spleen control group.

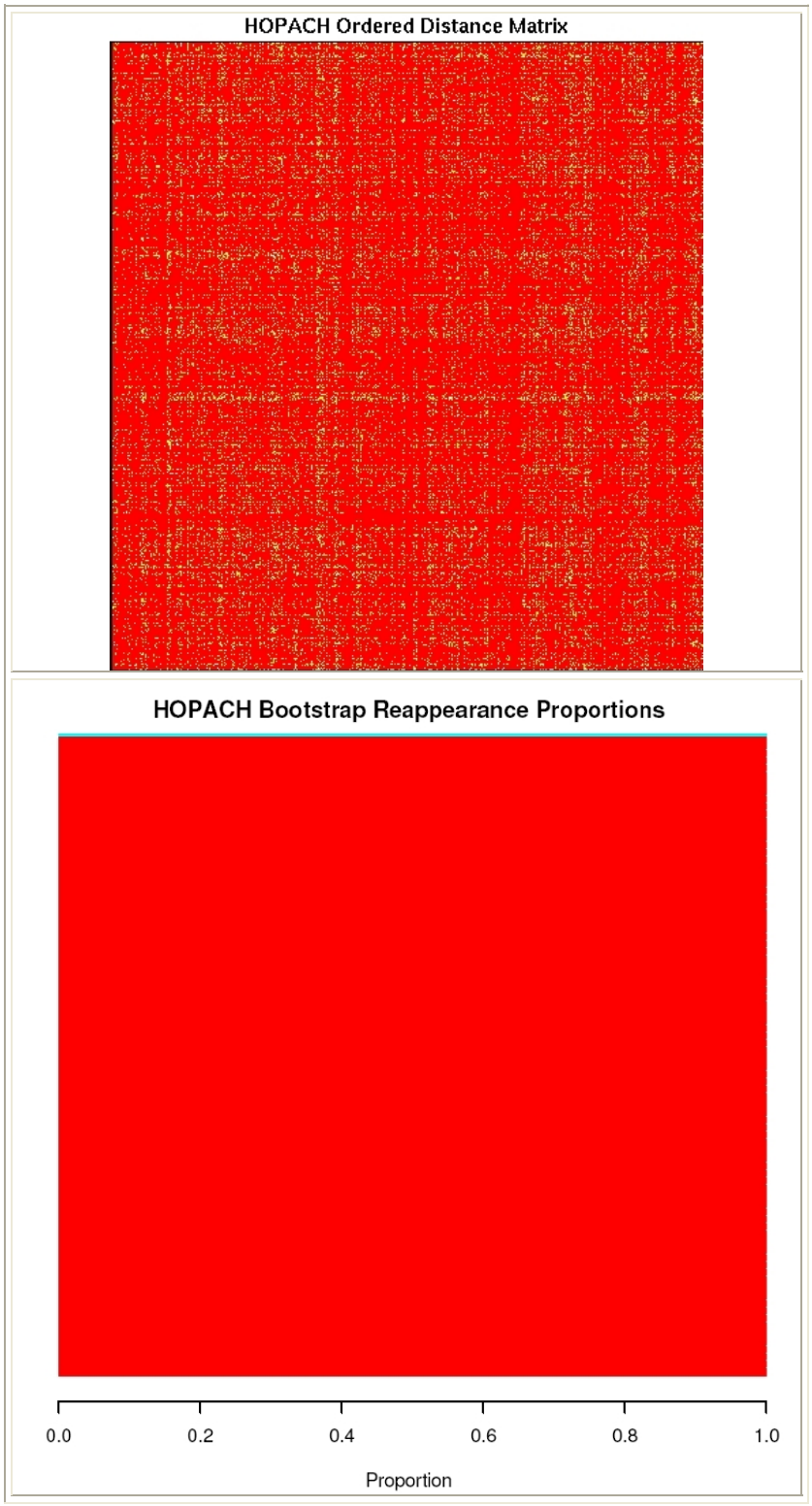


Figure 2.35: Ordered distance matrix and bootstrap reappearance proportions of genes that had a significant variance ratio in one and only one strain in the skin treatment group.

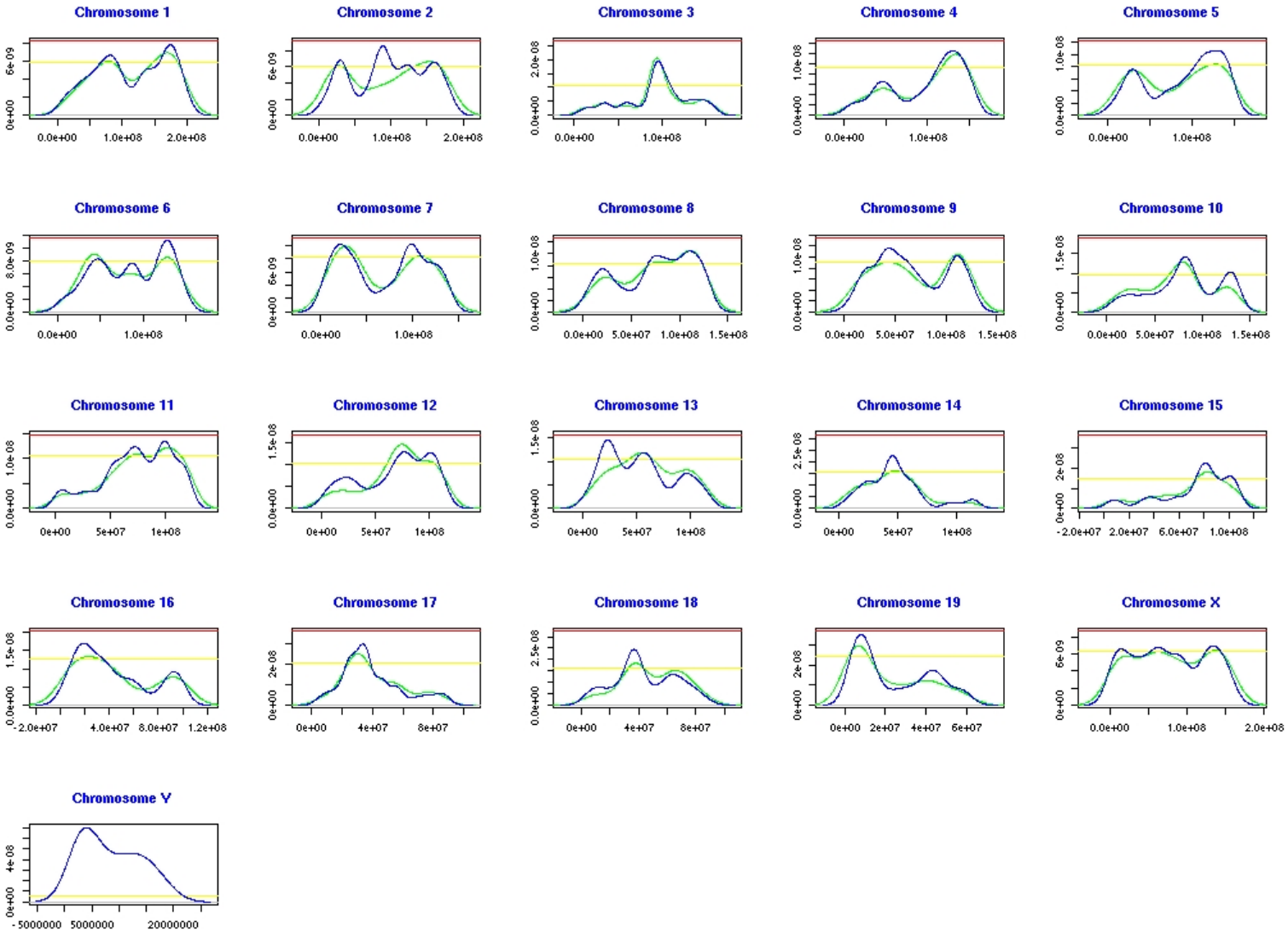


Figure 2.36: Gene density plots of genes that have a significant variance ratio in one and only one strain in the skin treatment group.

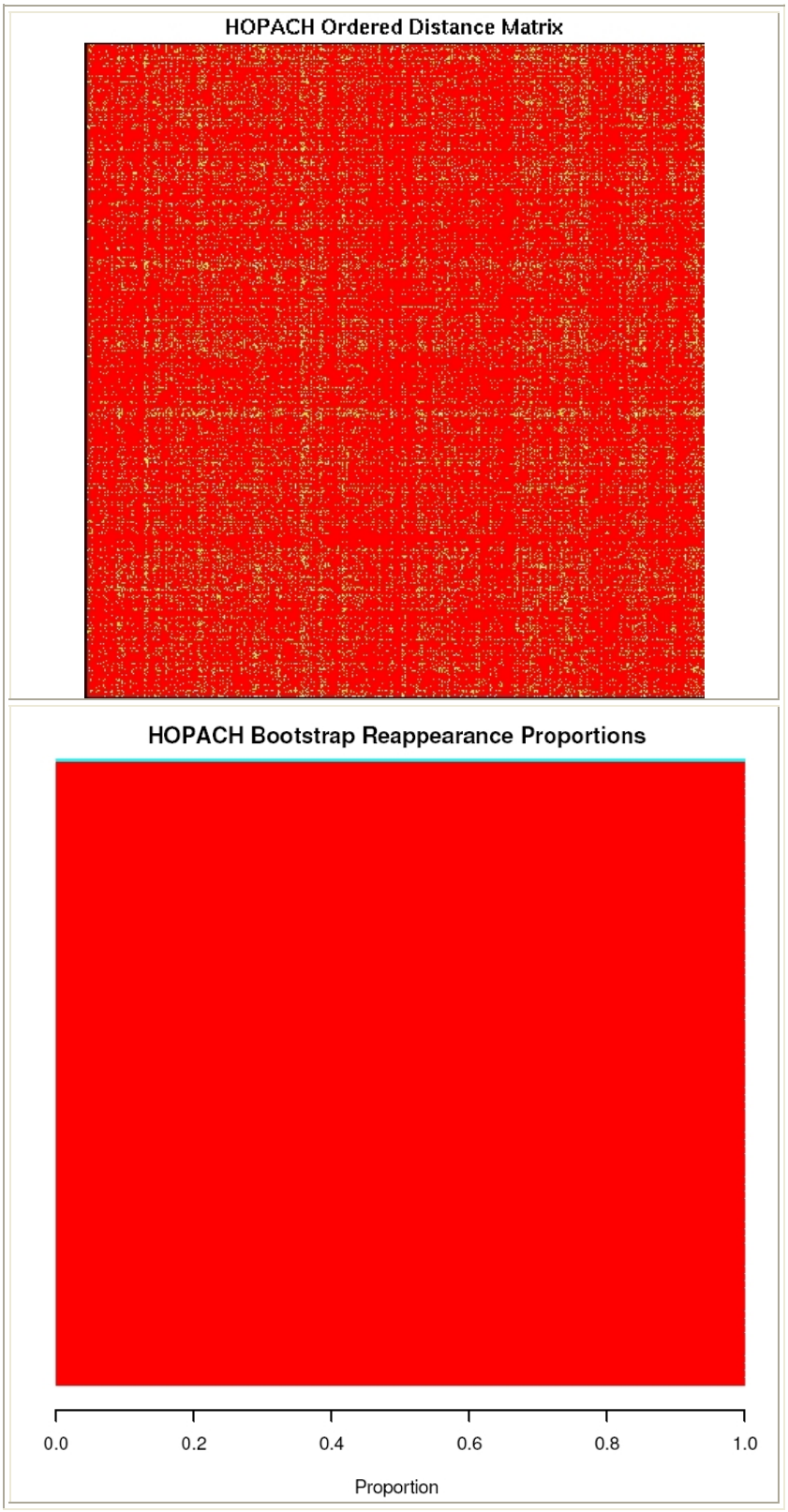


Figure 2.37: Ordered distance matrix and bootstrap reappearance proportions of genes that had a significant variance ratio in one and only one strain in the spleen treatment group.

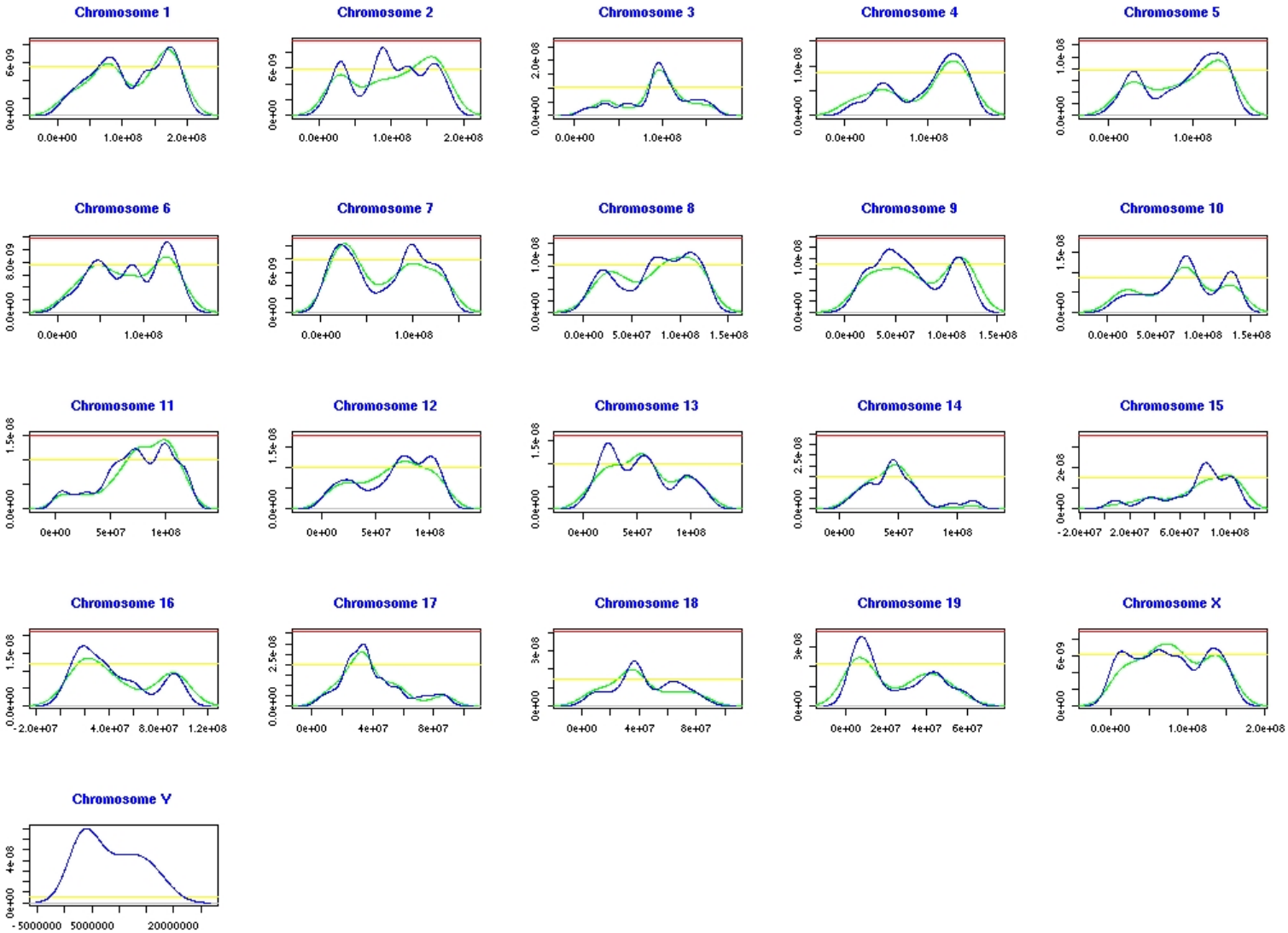


Figure 2.38: Gene density plots of genes that have a significant variance ratio in one and only one strain in the spleen treatment group.

also shows the barplot of the bootstrap reappearance proportions. From this plot, all of the genes belong entirely to the same cluster, so bootstrap had no effect on the data.

Since both of the clusters found by HOPACH had more than 10 genes in them, no new clusters needed to be created. There were 381genes in this set that were not annotated with Gene Ontology terms. One GO term was found to be enriched in this set: transcription factor complex (cellular components).

16. Genes that were significantly differentially expressed between the control and treatment groups in one and only one strain in the skin group. There were 1,857 genes in this set, 284 of which did not have a Gene ID. This reduced the set to 1,573 genes, 372 of which did not have an assigned chromosomal location. The gene density plots of each chromosome are in *Figure 2.40*. There were no gene-dense regions found on any of the chromosomes from this set.

Since this set was extremely large, clustering was performed using the HOPACH algorithm. There was only one variable available for clustering with this set, which was the t-test value for each significant gene in the strain. From this data, the HOPACH algorithm identified 2 main clusters in the 1st level of the tree, with a median split silhouette of -1.000. *Figure 2.39* shows the ordered distance matrix of the genes in this set, which shows that the two levels of the tree don't correlate with each. *Figure 2.39* also shows the barplot of the bootstrap reappearance proportions. From this plot, all of the genes belong entirely to only one of the two clusters.

Since both of the clusters found by HOPACH had more than 10 genes in them, no new clusters needed to be created. There were 0 genes in this set that were not annotated with Gene Ontology terms, and no terms were found to be enriched in either cluster.

17. Genes that were significantly differentially expressed between the control and treatment groups in one and only one strain in the spleen group. There were 886 genes in this set, 112 of which did not have a Gene ID. This reduced the set to 774 genes, 82 of which did not have an assigned chromosomal location. The gene density plots of each chromosome are in *Figure 2.41*. There was one gene-dense regions found on chromosome 16. The region under the peak of the density curve for this chromosome covered four megabases: from 0 to 40,000,000 bp. There were 20 genes from this set on chromosome 16, 16 of which were underneath the peak in the curve. The genes under the curve were: periplakin, ataxin 2 binding protein 1, protamine 2, G1 to phase transition 1, snail homolog 2 (Drosophila), RIKEN cDNA 4933404G15 gene, stromal cell-derived factor 2-like 1, synaptosomal-associated protein, DiGeorge syndrome critical region gene 6, DiGeorge syndrome critical region gene 8, polymerase (RNA) II (DNA directed) polypeptide H, histidine-rich glycoprotein, mucin 4, ribosomal protein L35a, zinc finger protein 148 and RIKEN cDNA 2010005H15 gene. These genes had start sites of 4575804, 6566524, 10362916, 10783139, 14243425, 16311915, 16760918, 17010508, 17657698, 17858887, 20341415, 22655909, 32447691, 32765514, 33113873 and 35866435 respectively. Only one of these genes was not annotated with a Gene Ontology category, but the remaining 15 genes were not found to be enriched for any GO categories.

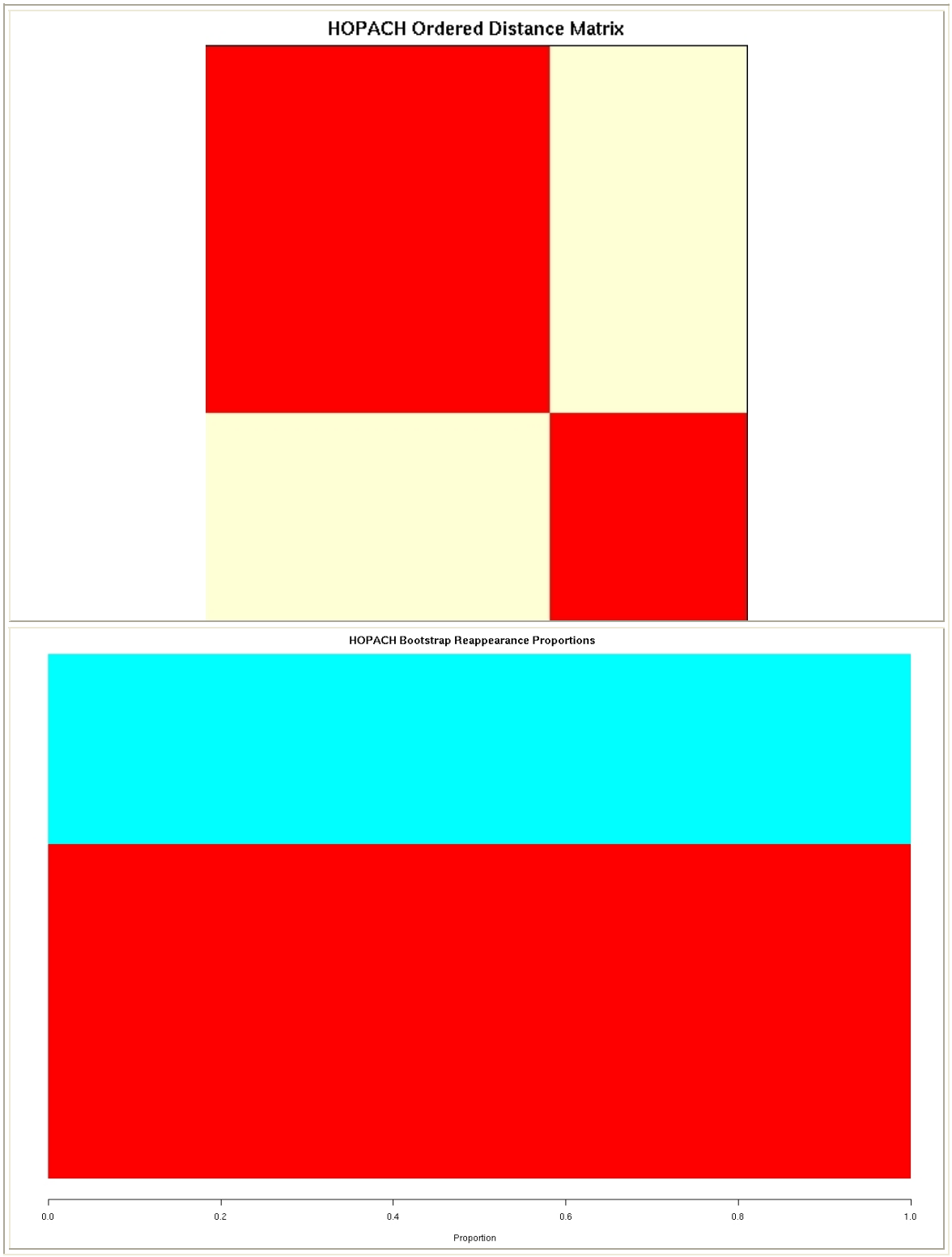


Figure 2.39: Ordered distance matrix and bootstrap reappearance proportions of genes that were significantly differentially expressed between the control and treatment groups in one and only one strain in the skin group.

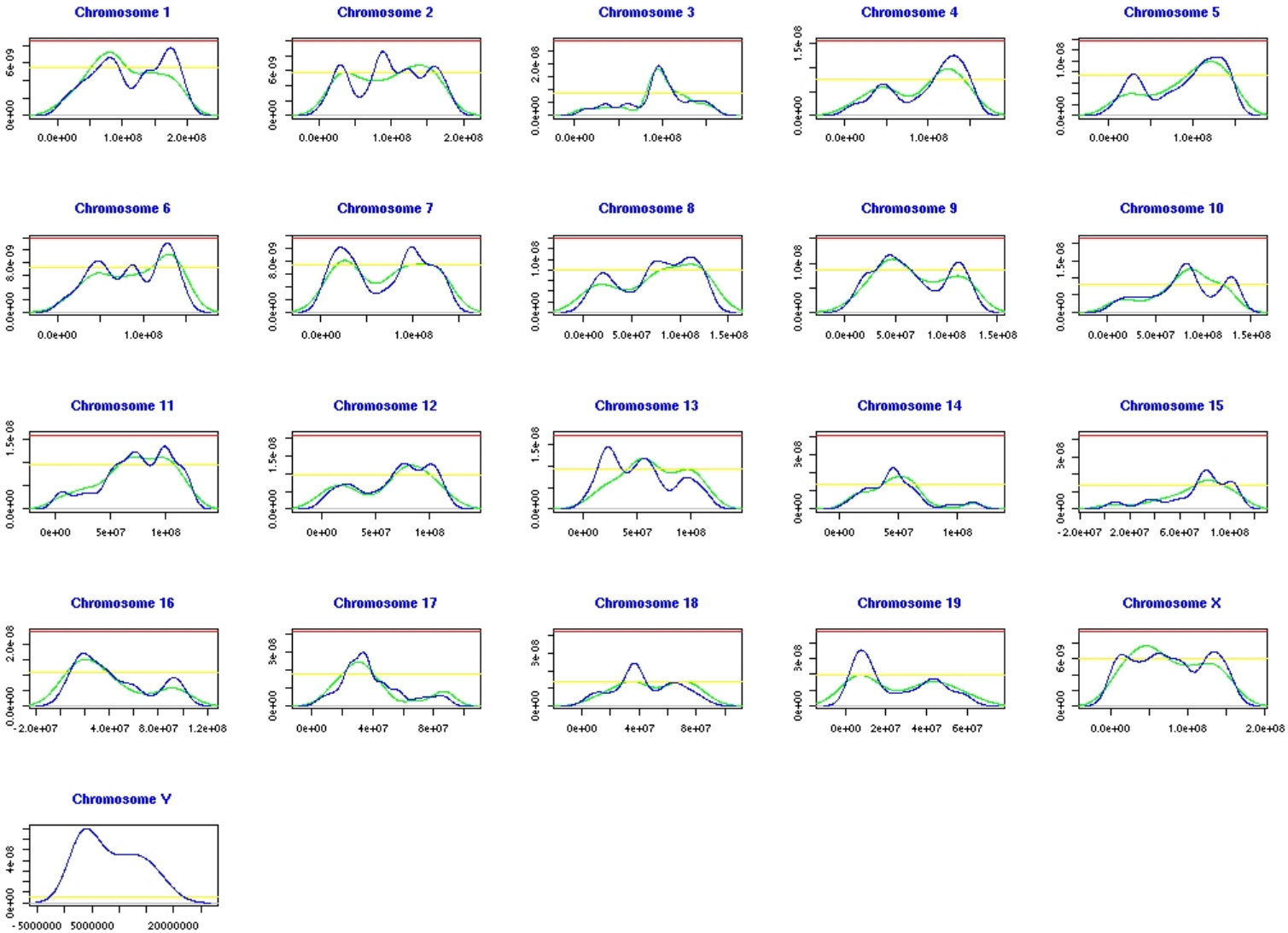


Figure 2.40: Gene density plots of genes that are significantly differentially expressed between the control and treatment groups in one and only one strain in the skin group.

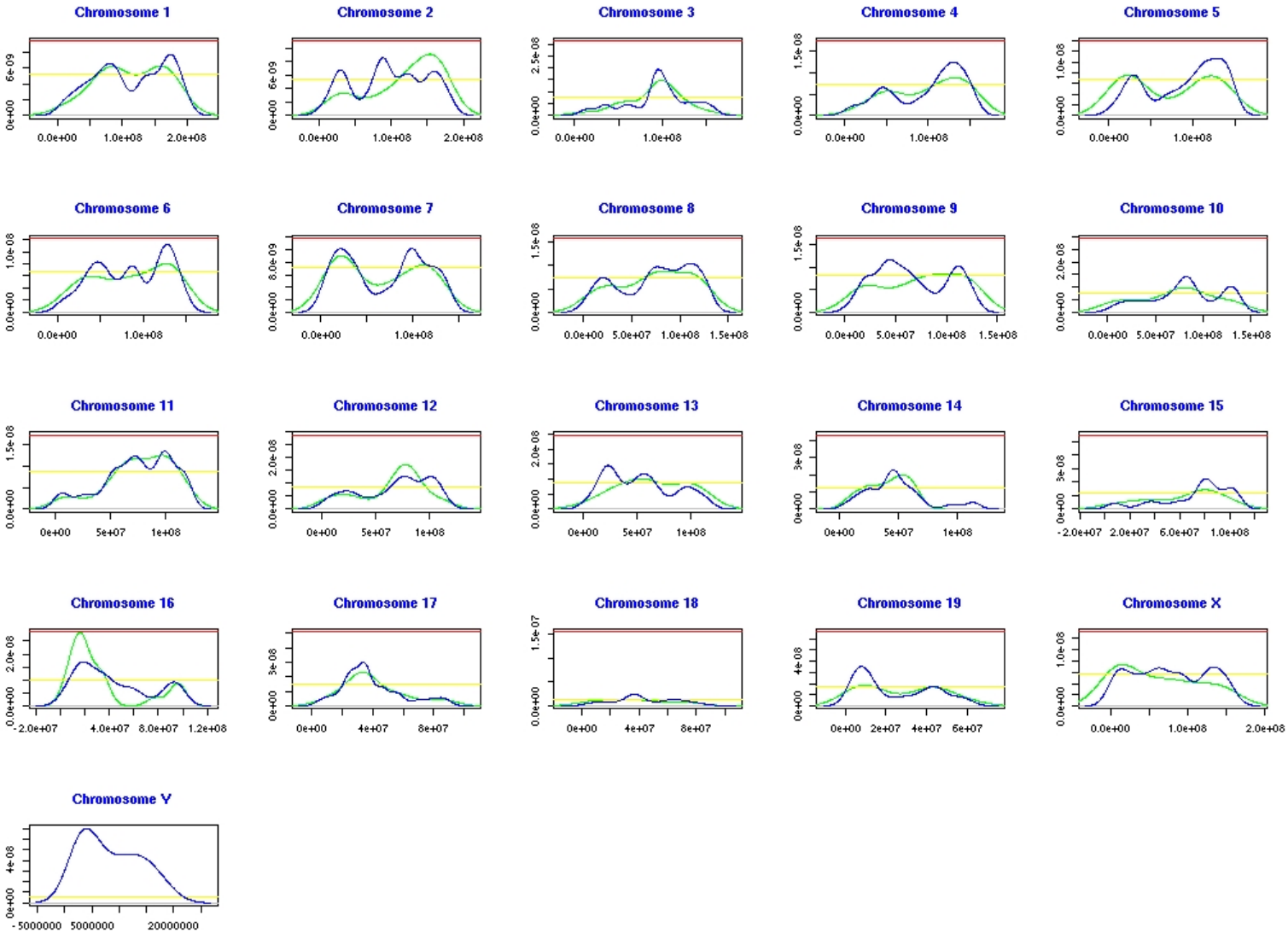


Figure 2.41: Gene density plots of genes that are significantly differentially expressed between the control and treatment groups in one and only one strain in the spleen group.

Since this set was extremely large, clustering was done using the HOPACH algorithm. There was only one variable available for clustering with this set, which was the t-test value for each significant gene in the strain. From this data, the HOPACH algorithm identified 2 main clusters in the 1st level of the tree, with a median split silhouette of -1.000. *Figure 2.42* shows the ordered distance matrix of the genes in this set, which shows that the two levels of the tree don't correlate with each. *Figure 2.42* also shows the barplot of the bootstrap reappearance proportions. From this plot, all of the genes belong entirely to only one of the two clusters.

Since both of the clusters found by HOPACH had more than 10 genes in them, no new clusters needed to be created. There were 0 genes in this set that were not annotated with Gene Ontology terms, and no terms were found to be enriched in either cluster.

18. Genes that were significantly differentially expressed between the treatment and control group, and had a significant variance ratio in both the control and treatment groups, in one and only one strain in the skin. There were 502 genes in this set, 42 of which didn't have a Gene ID. This reduced the set to 460 genes, 246 of which did not have an assigned chromosomal location. The gene density plots of each chromosome are in *Figure 2.44*. There was a region on chromosome 3 found to have a significant density of genes on it. There were 8 genes from this set on chromosome 3, 7 of which are under the peak in the density curve. The peak in the curve spans 2.5 megabases, and contains the following genes: RIKEN cDNA 6330415M09 gene, farnesyl diphosphate synthetase, RIKEN cDNA 4932431F02 gene, RIKEN cDNA 4932431F02 gene, RIKEN cDNA 1110058A15 gene, RIKEN cDNA 2310007A19 gene and RIKEN cDNA 2610022K04 gene. These genes had start sites of 88863530, 91299735, 92413789, 92413789, 95637164, 97162792 and 97835296 respectively. Four of these genes were not annotated with a Gene Ontology term, and there were no terms found to be enriched in the remaining 3 genes.

Since this set was so large, clustering was performed using the HOPACH algorithm. There were three variables used for clustering: the t-test value, the variance ratio for the control group and the variance ratio for the treatment group. From this data, the HOPACH algorithm identified 117 main clusters in the 5^{th} level of the tree, with a median split silhouette of 0.2781948. *Figure 2.43* shows the ordered distance matrix of the genes in this set. There are surprisingly few clusters that positively correlate with one another, as most of the clusters have no correlation with each other, as indicated by the yellow color. *Figure 2.43* also shows the barplot of the bootstrap reappearance proportions. Many of the genes belong to the first HOPACH cluster. Most of the genes belong to 5 different clusters, and none belong exclusively to one cluster. Almost all of the genes have equal bootstrap reappearance proportions in each cluster to which they belong, which is something not present in most of the other sets.

Since there were so many clusters derived by HOPACH, clusters with less than 10 genes in them were combined into clusters of 20 or more genes, as described in the Methods section. This yielded 68 clusters that had at least 10 genes each in them. From this set, there were 258 genes that did not have Gene Ontology annotations. Each cluster was analyzed for enriched terms, as described in the Methods section, but no terms were found to be enriched in this set.

The above results are summarized in *Table 2.7* as well.

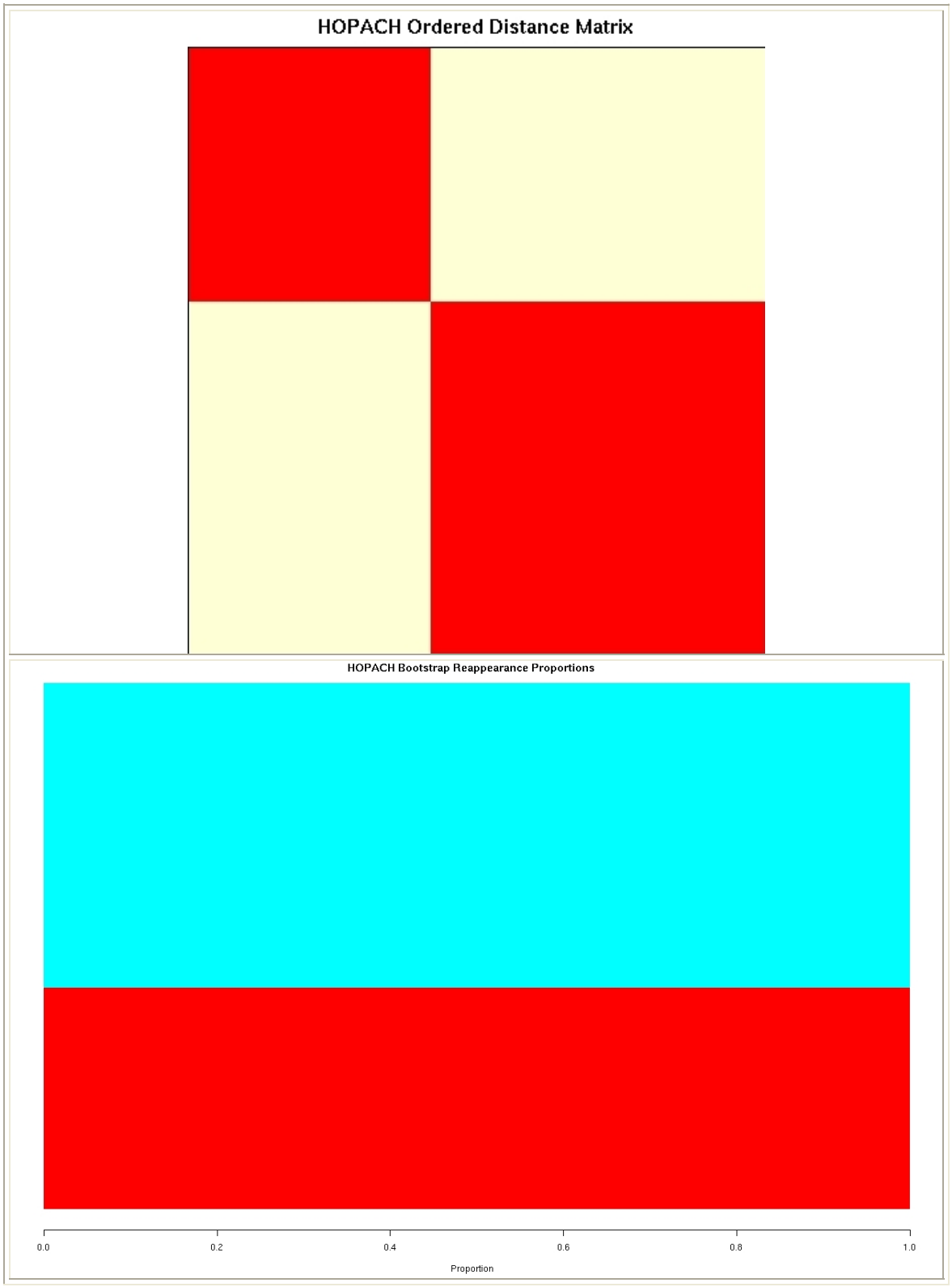


Figure 2.42: Ordered distance matrix and bootstrap reappearance proportions of genes that were significantly differentially expressed between the control and treatment groups in one and only one strain in the spleen group.

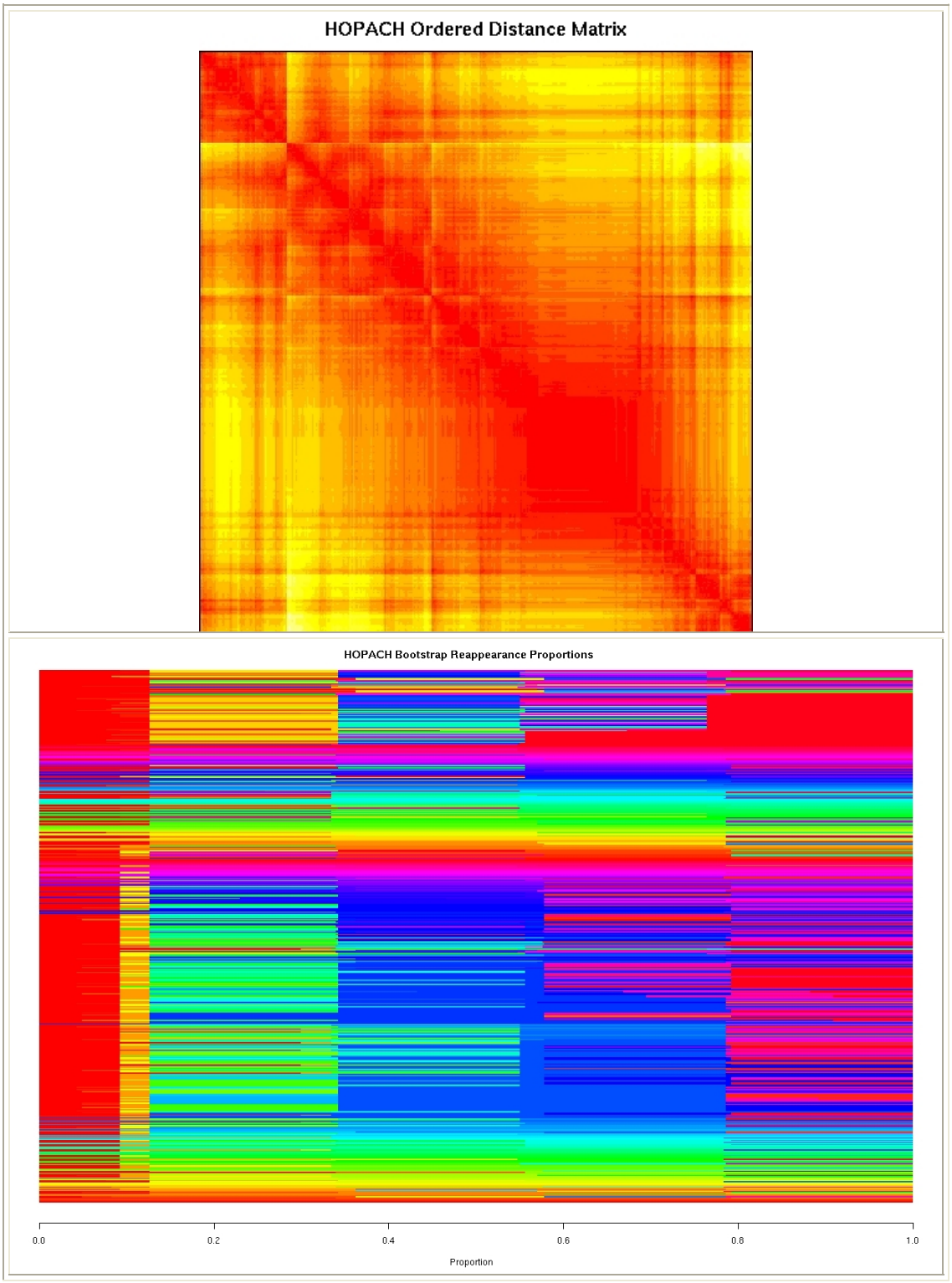


Figure 2.43: Ordered distance matrix and bootstrap reappearance proportions of genes that were significantly differentially expressed between the control and treatment groups, and had a significant variance ratio in both the control and treatment groups in one and only one strain in the skin.

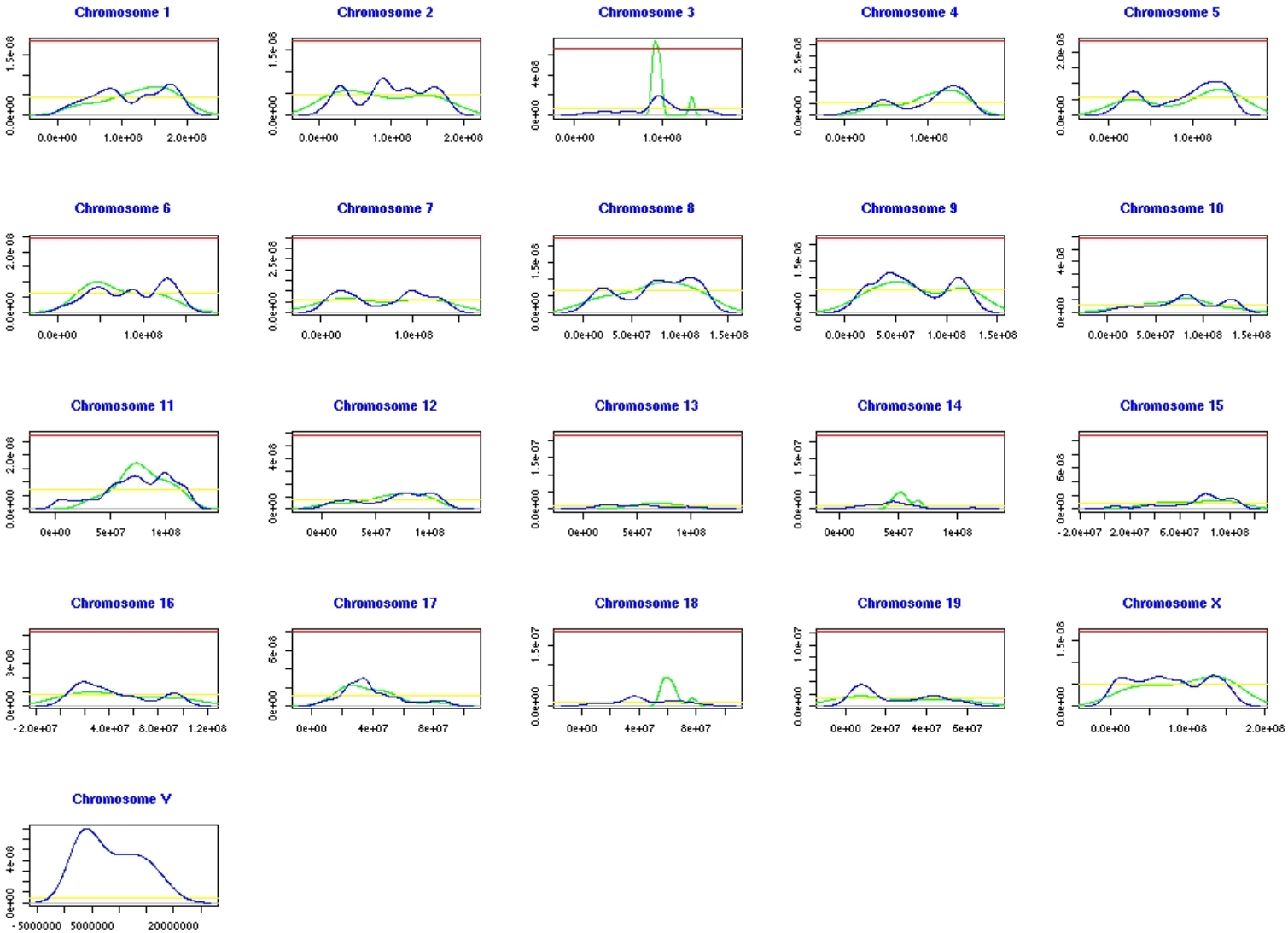


Figure 2.44: Gene density plots of genes significantly differentially expressed between the treatment and control group, and had a significant variance ratio in both the control and treatment groups, in one and only one strain in the skin.

Table 2.7: Summary of results from the analysis of different sets of genes.

Gene Set	Enriched Chromosome	Enriched GO Terms
Significant variance ratio in all 6 strains in the skin control group	None	None
Significant variance ratio in 5 strains in the spleen control group	None	None
Significant variance ratio in 5 strains in the skin treatment group	None	None
Significant variance ratio in 5 strains in the spleen treatment group	None	None
Significant variance between the strains but not within in the skin treatment group	None	None
Significant variance between the strains but not within in the spleen treatment group	None	None
Significant variance ratio in the treatment group, but not the control group, in 5 of the strains in the skin tissue	None	Male sex differentiation, integrin-mediated signaling pathway, external side of plasma membrane and cell surface
Significant variance ratio in the treatment group, but not the control group, in 5 strains in the spleen tissue	None	None
Significant variance ratio in the control and treatment groups, and differential expression between these groups in the skin tissue	Chromosome 3: 2.5 Mb Chromosome 11: 5 Mb	None
Differentially expressed between the control and treatment groups in all 6 strains in the skin tissue	None	Molecular function unknown
Differentially expressed between the control and treatment groups in all 6 strains in the spleen tissue	None	None
Significant variance ratio in only one strain in the skin control group	None	None
Significant variance ratio in only one strain in the spleen control group	None	None
Significant variance ratio in only one strain in the skin treatment group	None	None
Significant variance ratio in only one strain in the spleen treatment group	None	None
Differentially expressed between the control and treatment groups in only one strain in the skin treatment group	None	None
Differentially expressed between the control and treatment groups in only one strain in the spleen treatment group	Chromosome 16: 4 Mb	None
Differentially expressed between the treatment and control group, and significant variance ratio in both the control and treatment groups in only one strain in the skin	Chromosome 3: 2.5 Mb	None

Gene sets with too few genes to analyze:

The following sets of genes did not have enough genes in them to perform any type of quantitative analysis for the reasons given:

1. Genes with a significant variance ratio and were differentially expressed between strains in the skin control. There were 7 genes in the set, none of which had a Gene ID.

2. Genes with a significant variance ratio and were differentially expressed between strains in the spleen control. There were 0 genes in the set.

3. Genes with a significant variance ratio and were differentially expressed between strains in the skin treatment group. There were 0 genes in the set.

4. Genes with a significant variance ratio and were differentially expressed between strains in the spleen treatment group. There were 0 genes in the set.

5. Genes with a significant variance ratio in the control and treatment groups, and was differentially expressed between these groups in the spleen tissue. There were 0 genes in the set.

6. Genes that were significantly differentially expressed between the treatment and control group, and had a significant variance ratio in both the control and treatment groups, in one and only one strain in the spleen. There were 0 genes in this set.

Pooled Data:

The same tests were applied to this second data set, as was applied to the above data. The data was normalized using the procedure described in the Methods section, where the treatment group was the cDNA samples that came from a pool, and the control group was the cDNA samples that came from a single animal.

There were hundreds of genes found to be differentially expressed (FDR $p < .001$) between the single-animal and pooled groups in the skin tissue (*Figure 2.45*). There was only data from a single strain available in the spleen tissue, and almost no genes were found to be differentially expressed. Because of this, variance ratios could only be calculated in the skin tissue. As shown in *Figure 2.46*, there was a large increase in the number of genes with a significant variance ratio ($p < .01$) in the pooled-animal group as opposed to the single-animal group. The same basic trends are seen with this data, as with the above data set when comparing the $REML_{between}$ and $REML_{within}$ estimates (*Figures 2.47 and 2.48*). The single-animal group had much smaller $REML_{between}$ and $REML_{within}$ estimates than did the pooled group. As before, there were some extreme outliers in *Figure 2.48*, and these were removed from the plot.

Figures 2.49 and 2.50 show the gene density plots with the genes that have a large variance estimate within a strain, but not between.

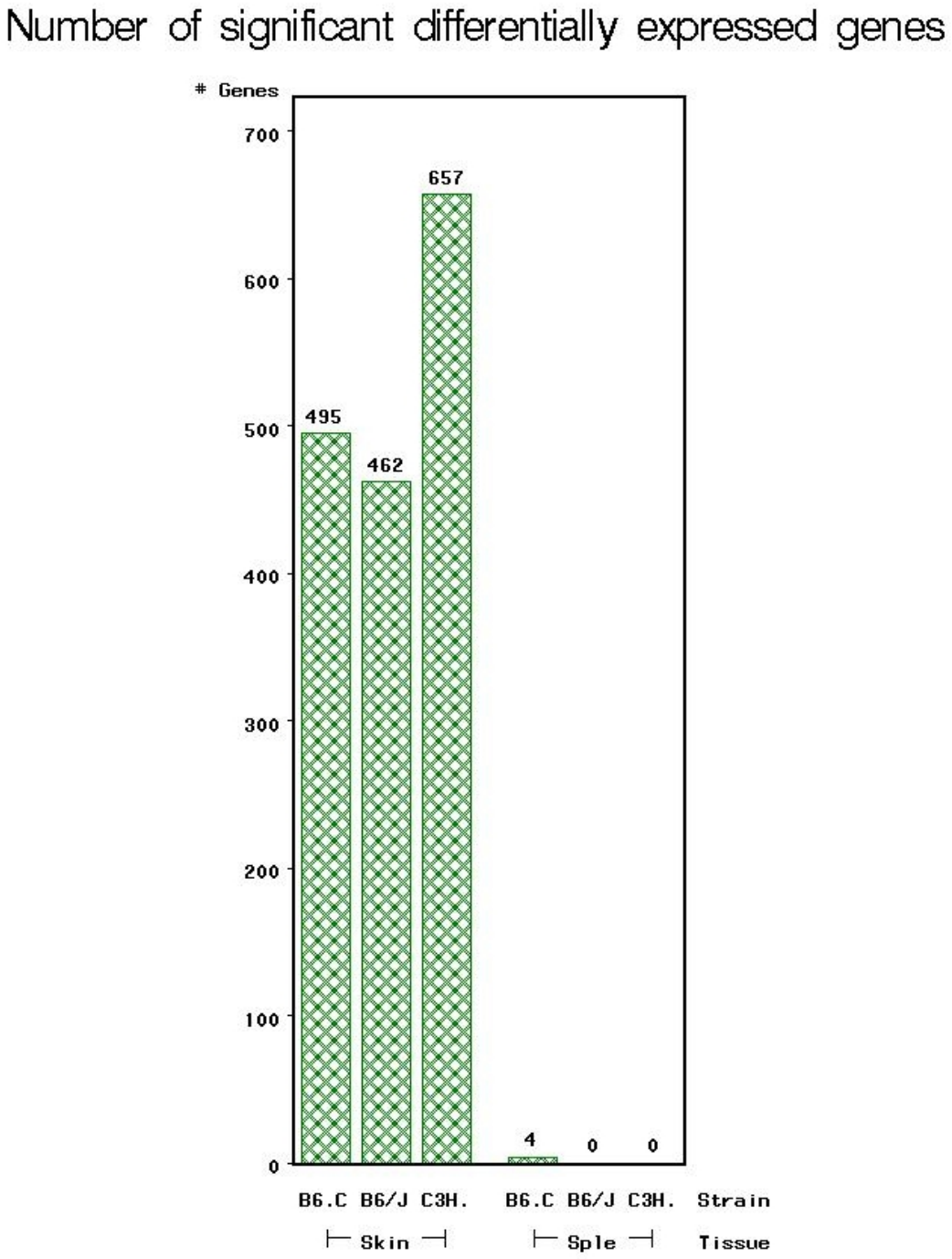


Figure 2.45: The number of significantly differentially expressed genes between the single and pooled samples of RNA in each tissue and strain. The spleen tissue had only the B6.C, so B6/J and C3H. can be ignored.

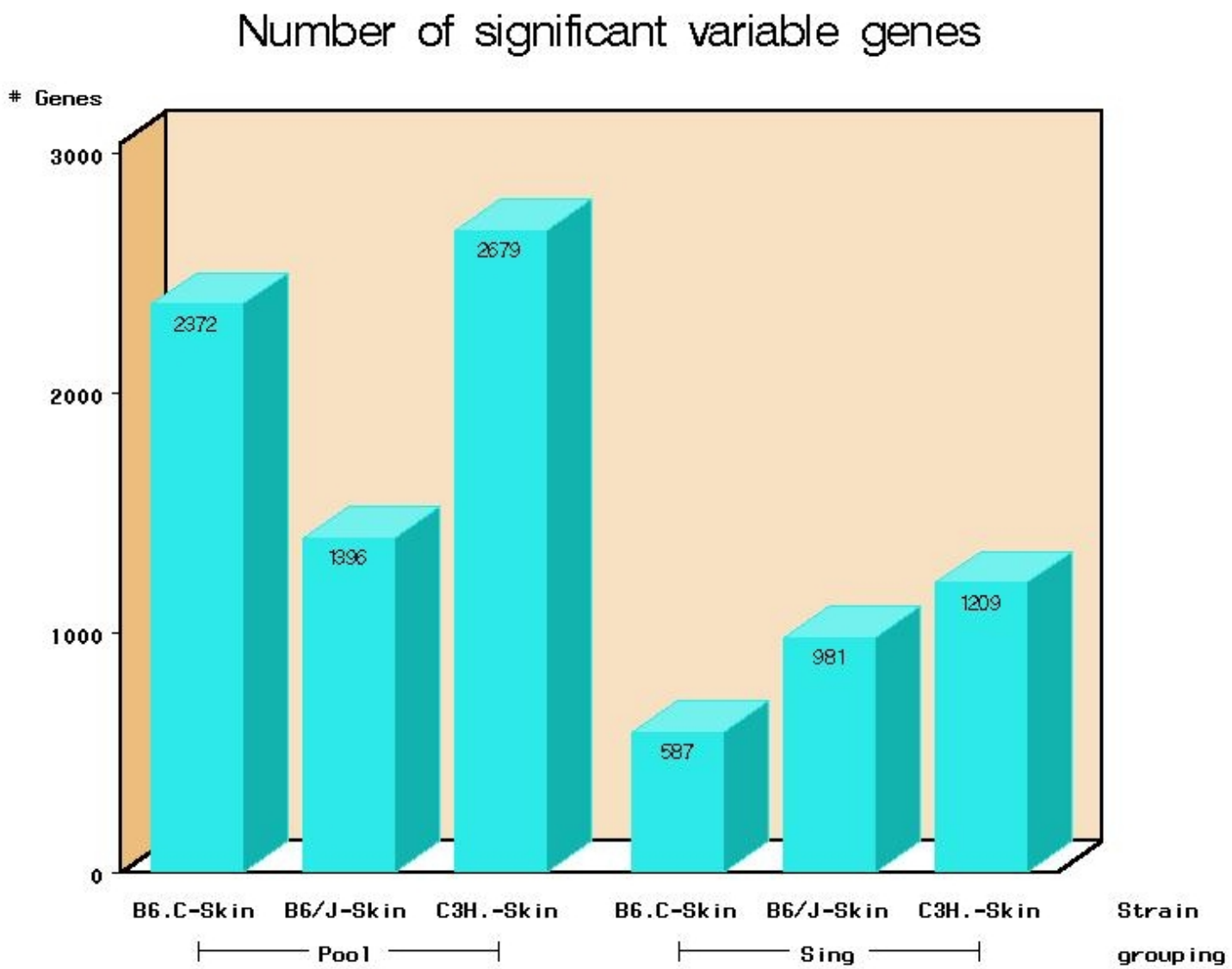


Figure 2.46: Number of genes in each strain with a statistically significant variance ratio ($p < .05$).

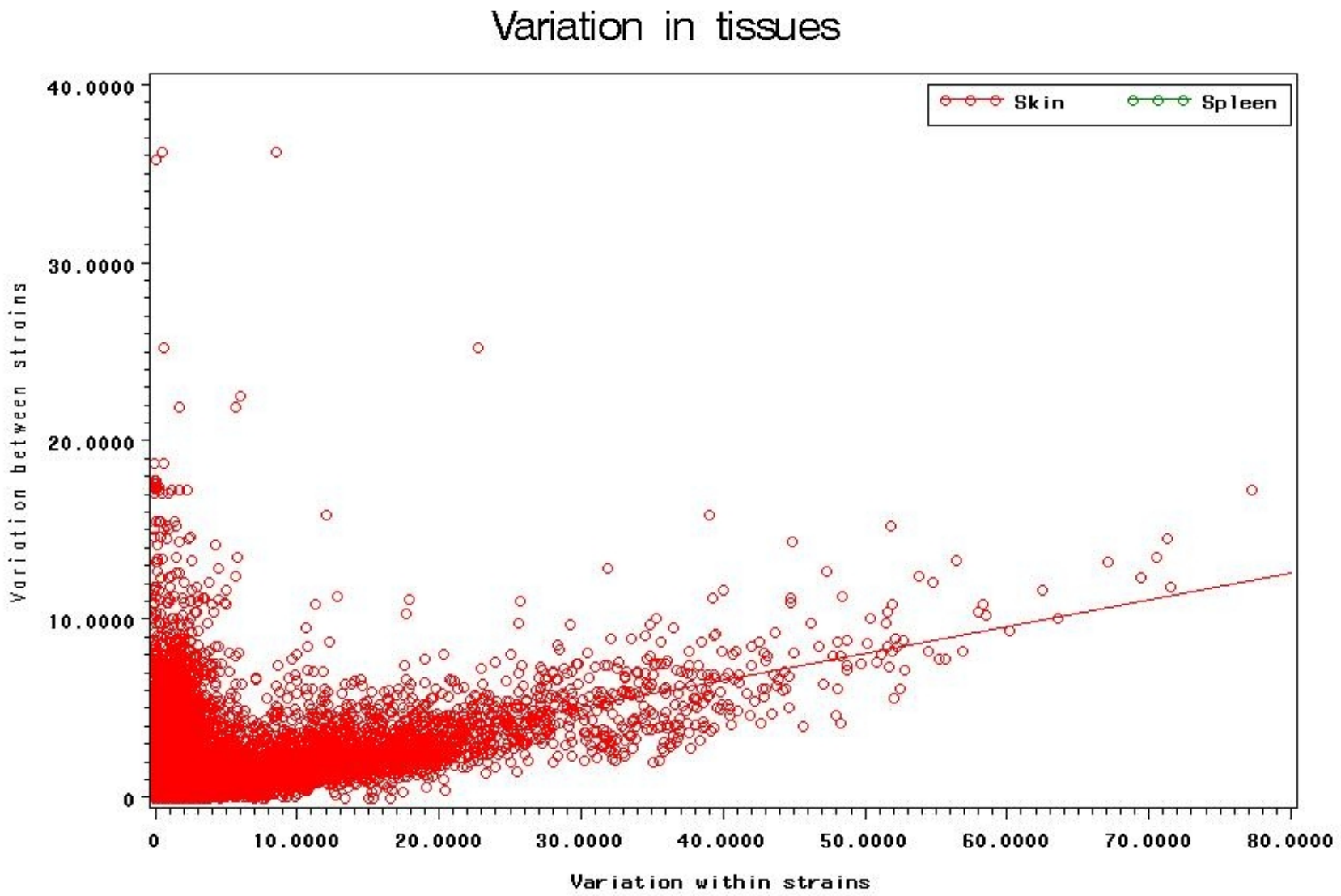


Figure 2.47: REML estimate between strains vs. within strain for the single-animal RNA. Red points represent skin.

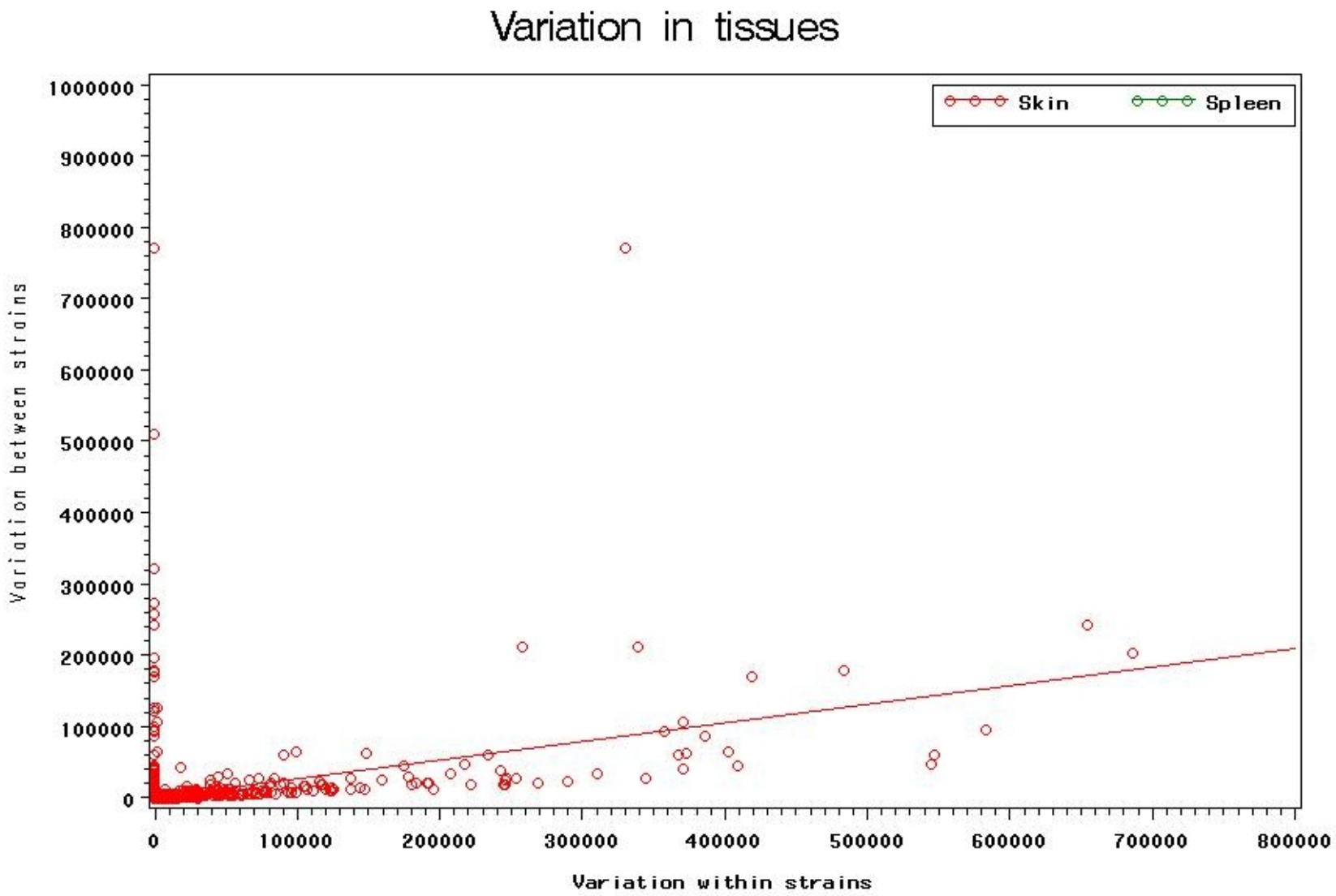


Figure 2.48: REML estimate between strains vs. within strain for the pooled RNA. Red points represent skin. Several extreme outliers were removed from this plot.

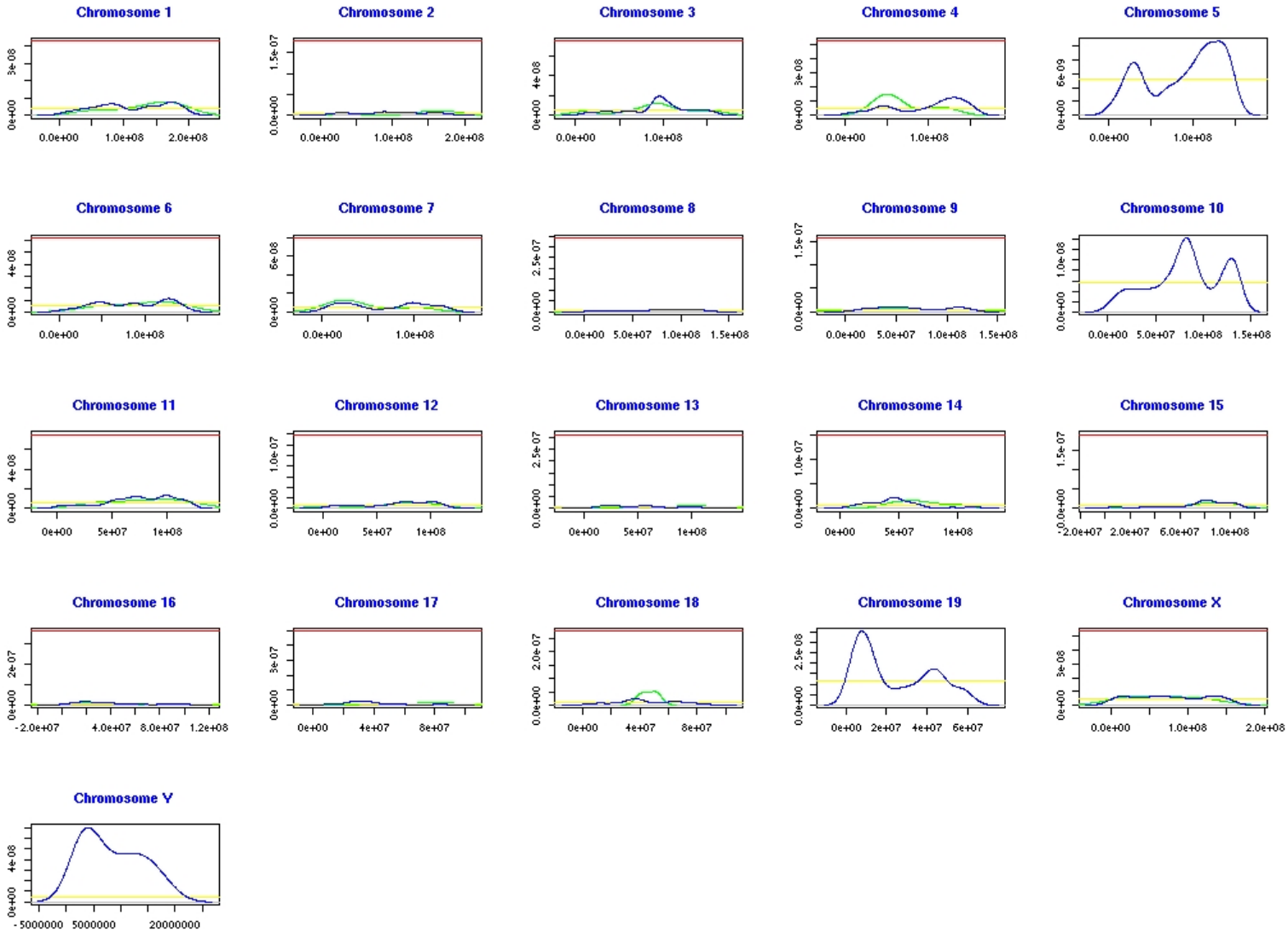


Figure 2.49: Gene density plots of genes with a significant variance ratio within the strains, but not between in the skin treatment.

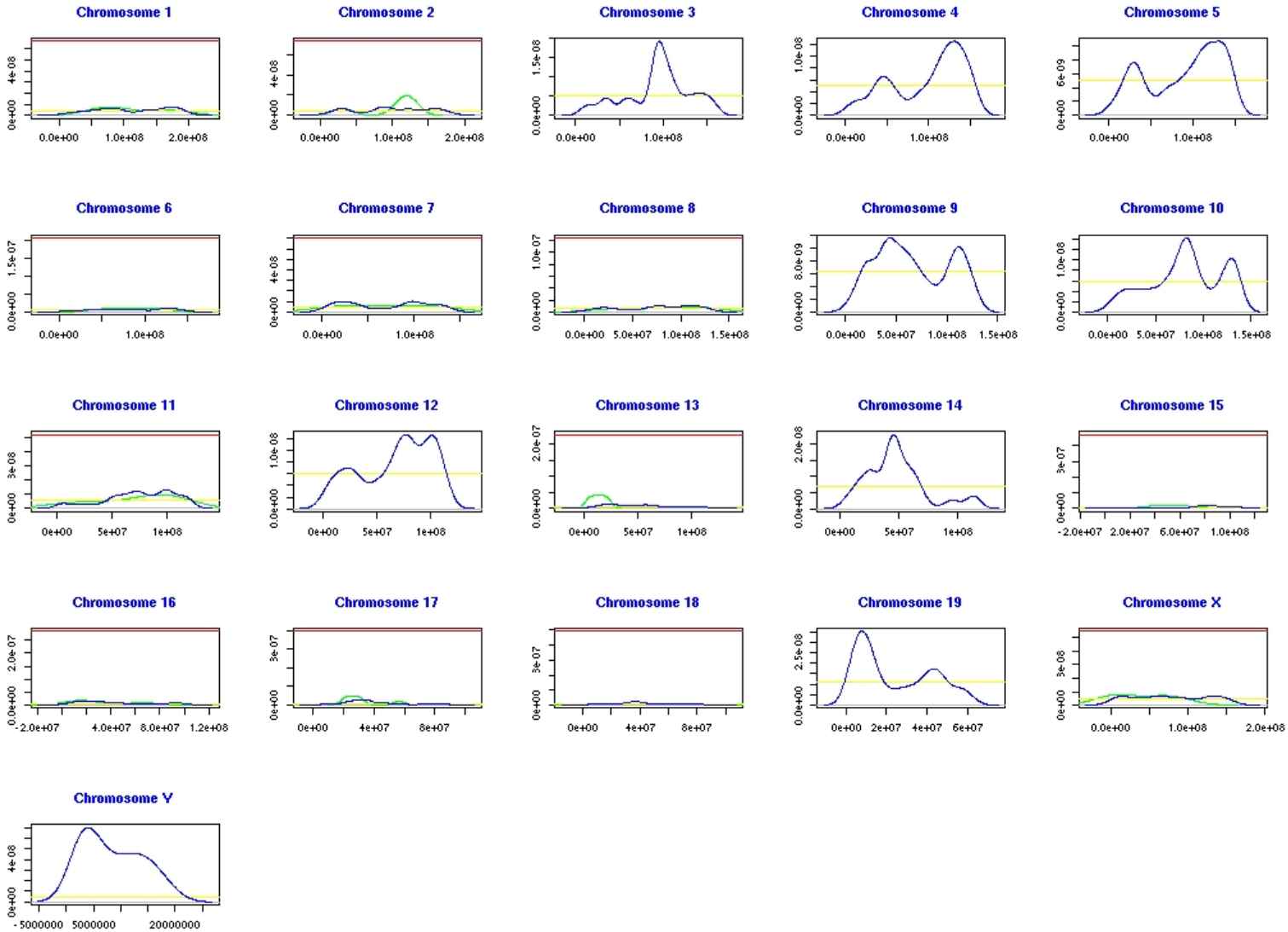


Figure 2.50: Gene density plots of genes with a significant variance ratio within the strains, but not between in the spleen treatment.

V. Discussion

The primary focus of this study was to identify inter-individual differences in mice, in response to low-dose radiation treatment. On the whole, radiation treatment had a dramatic effect on a mouse's gene expression, both in terms of variability and differing expression. In both tissues and every strain, there was an increase in the number of genes that had a significant variance ratio. The radiation treatment also appears to have had a more dramatic effect on the skin tissue, than the spleen tissue, as there are many more genes with a significant variance ratio. The estimates for the between- and within-strain variance were also dramatically larger in the treatment group, as compared to the control. This is evident by the scales of the plots in *Figures 2.4 and 2.5*. It is possible that such large variances were obtained in the treatment group because the sample size was too small [95]. The trends in these plots are of interest as well, in that *Figure 2.4* has points in both tissues where there is an increasing amount of between-strain variance with very little within-strain variance. This then winds into a linear trend between the two variables, but very few points are present with a large within-strain variance and small between-strain variance. This is not the case in *Figure 2.5* though, where there are still these same two trends, but also a large number of genes that have a large within-strain variance but small between-strain variance. This shows that animals with the same genetic background respond differently to the radiation treatment. Since these animals have the same genetic constitution, the only factor that could be contributing to these differences in gene expression could be environmental influences. Such factors could include litter size, the number of male versus female pups in a litter, or the age of the mom. The large between-strain variance is due to the fact that inbred strains are known to differ slightly genetically due to the presence of single nucleotide polymorphisms (SNPs). It has been shown that in inbred mice strains, there are blocks of SNPs in genes located throughout the genome [33], which can easily lend to the variation in gene expression noticed here. Because of this, identifying the chromosomal location of highly variable genes is of interest. As shown in *Figures 2.49 and 2.50* though, genes with a large within-strain variance but small between-strain variance don't have any gene-dense clusters on any of the chromosomes in either the skin or spleen tissue. It would be expected that if blocks of SNPs were playing an important role in the animal's response to radiation, that there would be genes with a large variance clustered together in the genome.

In addition to empirically looking at the differences between the control and treatment groups by assessing differences in the variance ratio, multiple pair-wise comparisons were made between these groups. Even with a false discovery rate (FDR) correction, many genes were found to be differentially expressed in both tissues and all strains. As with the variance ratios, the skin had a larger number of significant genes than did the spleen tissue. The B6.C and Bl6Jx strains had almost twice as many genes differentially expressed than any of the other strains in the skin tissue. They had also had many more genes with a significant variance ratio in the skin treatment group, although there wasn't as large of a disparity between these and the other strains. There was no different between these two strains and the other four strains in the spleen tissue with either test though.

Another test for differential expression was also done between the different strains in each tissue and treatment type. This attempted to assess inter-strain differences in the means of the residuals of the expression values. No multiple comparison adjustment was made with this test, since so few genes were found to be significant. While more genes were found to be differentially expressed in the control group compared to the treatment group, the numbers are so small that this cannot be interpreted as being significant. Therefore, it can be assumed that in comparison of the means of the strains, there are no significant differences in expression.

Since this study was focused around the characterization of inter-individual differences in low-dose radiation exposure response, it was important to know inter-individual similarities as well. This was accomplished by identifying genes that were variable and differentially expressed in more than one strain. These genes would be expected to be involved in typical response pathways, such as DNA damage response, or common "housekeeping" pathways, such as the cell cycle. Surprisingly though, as shown in *Table 2.3*, there are no genes in the either tissue of the treatment group that have a significant variance ratio in all six strains. There are many genes common to 5 of the 6 strains in both tissues, and many more in the treatment groups than in the control groups. It is expected that in the control groups there are few genes with a significant variance ratio in most of the strains, since common housekeeping genes are not expected to have dramatic differences in their expression levels. There are 10 genes in the skin control group that have a significant variance ratio in all 6 strains though. It would be expected that there would be genes with a significant variance ratio in the treatment groups though, as all mice have the same pathways, and it would be expected that the pathways relevant to radiation response would be activated as a result of the radiation exposure, which would increase variability of expression of the constituent genes. This however, was not the case. If anything this again points to inter-individual differences in response to low-dose radiation treatment.

A similar approach was taken to identify genes with a significant variance ratio in both tissues and between different strains. It would be expected that both the skin and spleen tissues might have some common functions in response to radiation exposure, so would have some of the same genes being actively transcribed. Interestingly enough though, there were no genes with a significant variance ratio in both tissues that were common to even 5 of the 6 strains. This was also true in the controls. There were some genes in the treatment group that were common between 4 of the 6 strains, as shown in *Table 2.4*.

This same interest was pursued with genes found to be differentially expressed between strains, although since so few genes were found to be significant, no conclusions can be drawn from this. In the test of differential expression between control and treatment groups, many genes were found to be significant in all of the inbred strains. There were 1,563 genes in the skin tissue, and 654 genes in the spleen tissue that were differentially expressed in all six strains. It is possible that many of these genes are involved in cellular responses to radiation treatment, or associated pathways.

A related critical question to consider was how many of these genes that were significantly variable or differentially expressed were significant in one and only one strain? Many genes were found to have a significant variance ratio in only one strain, as

shown in *Figure 2.9*. As before, there are more variable genes in the treatment group than the control, although BALBxJ in the skin, and B6.C and Bl6Jx in the spleen have almost the same number of genes in the treatment as the control. This is interesting, since it was these last two strains that had so many more genes differentially expressed in the skin between the control and treatment groups. These two strains seem to have this same effect (as do others) in the skin tissue, where there are more than twice as many genes variable between the control and treatment groups, but it is interesting that there are such dramatic differences in the two tissues. One possible interpretation of this is that the differing genetic backgrounds of the inbred strains are not contributing to the animal's response to radiation in the spleen tissue, but it is having considerable effects in the skin tissue.

Genes that were differentially expressed between the treatment and control groups were also analyzed for how many were significant in one and only one strain. B6.C had a considerable number of genes that were differentially expressed in the skin, which was more than four times as many genes as any of the other strains. The spleen tissue in each strain had a small number of genes differentially expressed in only one strain, with the exception of A/J, which had almost the same number of genes in both tissues. On the whole though, most of the strains did not have very many genes that were significantly differentially expressed in only one strain.

Another comparison identified how many genes were differentially expressed between the treatment and control groups in only one strain, and how many of these also had a significant variance ratio. Surprisingly, there were no genes in any of the strains in the spleen tissue that were both differentially expressed and variable in only one strain. In the skin tissue, B6.C and Bl6Jx had more than two hundred genes each (283 and 241 respectively), while the other four strains each had less than thirty. Since there are such a small number of genes with overlap between these two tests, it is possible that many genes are simply a statistical artifact, and not biologically meaningful. This could very well be the case with genes that have a large variance ratio, but are not differentially expressed. On the other hand, genes with a high variance ratio and differential expression are ones with only a few individuals changing expression, but apparently have a large enough change to yield a significant amount of differential expression.

Since these tests revealed some similarities between the strains, hierarchical clustering was performed on the arrays, to attempt to flush out any relationships between the strains. As shown in Figure *2.12*, the strains do not seem to cluster together, and there don't appear to be any consistent relationships between them. The patterns and distances (Euclidean distances were used) are not similar between the skin and spleen in either the control or treatment groups. In Figure *2.13*, the replicate arrays within a strain were averaged, and clustering was again performed. It was expected that there would be some consistent patterns of clustering between the strains, but this again was not the case. None of the strains cluster together in the same pattern between tissues or treatment groups. This could be affected by how many genes are variable in their expression, and the fact that most of these variable genes are different between the two tissues.

While taking an empirical view of the results of the different statistical tests performed is informative, it is also of interest to find biological significance in the genes identified. Since little is known about specific radiation exposure-related responses, and

how these change between individuals with differing genetic constitutions, it is of particular interest to identify genes that could potentially be involved in related pathways. There were two basic mechanisms for doing this: to look at regions of the chromosomes that had a disproportionate number of genes located near each other, and to identify any Gene Ontology categories that were enriched. There were 23 different sets produced, each of which is described in the Results section. Sets with more than 30 genes were clustered, to attempt to produce smaller groups of related genes. In this way, it would be more likely to identify groups enriched for Gene Ontology categories. Clustering appeared to have a positive effect on all of these sets, except for those for which clustering was performed on a single value (sets with genes unique to a single strain). In these sets, clustering only found two groups in each set, and little significance could be drawn from this.

Each cluster of each set was analyzed for enriched Gene Ontology categories, although disappointingly few were found. There were three enriched categories found in the genes that had a significant variance ratio in 5 of the 6 strains in the skin treatment group. These were: cytokine and chemokine-mediated signaling pathway, heparin binding and integrin-mediated signaling pathway. Each of these three categories had three genes each in them. Cytokine and chemokine-mediated signaling pathway was comprised of protein tyrosine phosphatase non-receptor type 6, Kruppel-like factor 6, and kit oncogene. The heparin binding category contained laminin gamma 2, serine (or cysteine) proteinase inhibitor, clade D, member 1 and cysteine rich protein 61. The third term, integrin-mediated signaling pathway contained integrin alpha 2, PTK2 protein tyrosine kinase 2 and a disintegrin and metalloprotease domain 22. In the set of genes that had a significant variance ration in 5 of the 6 strains in the treatment but not control groups in the skin tissue, and four terms were found to be enriched: male sex differentiation, integrin-mediated signaling pathway, external side of plasma membrane and cell surface. Each of these four categories also only had three terms in each of them. Male sex differentiation had androgen receptor, bone morphogenetic protein receptor, type 1A and platelet derived growth factor receptor, alpha polypeptide. Integrin-mediated signaling pathway had the same terms as listed before, and external side of plasma membrane contained thymus cell antigen 1, theta, CD200 receptor 1 and EGF-like module containing, mucin-like, hormone receptor-like sequence 1. The fourth term, cell surface contained the same genes as external side of plasma membrane. One term was enriched in the set of genes that were differentially expressed between the control and treatment groups in all six strains, but this term was molecular function unknown, which is not at all helpful. Seven genes from the set were listed in this category: RIKEN cDNA 5730494M16 gene, dynein 2 light intermediate chain, CDK5 regulatory subunit associated protein 3, ubiquilin 4, BCRABL transformation potentiating sequence, RIKEN cDNA 2810429O05 gene and excision repaiross-complementing rodent repair deficiency, complementation group 8. There was also one term found in the set of genes that had a significant variance ratio in only one strain in the spleen treatment group. This was surprising, given the fact that clustering of this set was ineffective. The enriched term found was transcription factor complex. 74 genes from the set were annotated with this term, and are too many to list here.

In addition to the identification of enriched Gene Ontology categories, each of the 17 sets large enough to be analyzed was plotted in gene density distributions along the chromosomes. A background distribution was created from the genome, and each set was used to create a second distribution, as described in the Methods section. In the set of genes that were differentially expressed between the control and treatment groups, and were also variable in both groups in the skin tissue, there were regions of interest identified on chromosomes 3 and 11. There were 12 genes from this set located on chromosome 3, 9 of which were under the peak in the density curve. Chromosome 11 had 22 genes from the set on it, 12 of which were underneath the peak in the density curve. Chromosome 16 was also found to have a gene-dense region in the set of genes that were significantly differentially expressed between the control and treatment groups in one and only one strain in the spleen tissue. There were 20 genes in this set located on chromosome 16, 16 of which were under the peak of the density curve. There was also a region of interest found on Chromosome 3 in the set of genes that were differentially expressed between the control and treatment groups, and had a significant variance ratio in both groups, in only one strain in the skin tissue. Eight genes from this set were found on chromosome 3, 7 of which were located under the peak of the density curve. Each of these regions was tested for enriched Gene Ontology categories, although none were found. The specific genes in each of these groups just mentioned are detailed in the Results section. There does not seem to be any obvious relationships between the genes in these dense regions, especially since most of these are Riken clones, for which there are few details available.

The final part of this study consisted of assessing changes in the variability of gene expression in samples that came from a single animal, and those that came from a pool of four animals. There was a large increase in the number of genes found to have a significant variance ratio in the skin tissue in the single animal samples compared to the pools. This increase wasn't as dramatic in the B6/J strain as it was in the B6.C and C3H strains, but there was still a large increase in the number of variable genes. The same trends were noticed in the relationship of between- and within-strain variances as in the low-dose study. In the single animal plot (*Figure 2.47*), genes had the same trend as the control group in the radiation study (*Figure 2.4*), where there were no genes with a large within-strain variance and small between-strain variance. In the plot of samples from the four-animal pools (*Figure 2.48*) though, this trend changes to look more like *Figure 2.5* (radiation treatment). There are many genes that have an increasing within-strain variation and small between-strain variance. This clearly shows that by using a tissue sample for a microarray hybridization that comes from a pool of animals, variability is being added into the experiment, which cannot be accounted for through statistical adjustment. In addition, several hundred genes were also found to be differentially expressed between the single animal and pooled samples (*Figure 2.45*), which may be statistical artifacts.

Chapter 3

Bayesian network analysis

I. Abstract

Pathway analysis and network construction methods have become a popular topic of interest in the bioinformatics community, which has been a result of the large volumes of data being produced through high-throughput genomics experiments. One popular method for building such pathways are through the use of Bayesian networks. Many questions remain unresolved in their application though, such as how many experiments are required before an accurate model can be produced, and how accurate is that model. Based upon the simulations run, a minimum of 300 data points is required to obtain a graph that resembles the true structure, while more than 300 data points continues to improve the accuracy. The same number of data points also consistently produced the highest scoring graph, and allowed for the identification of modules within the graphs. The accuracy of this decreased though as the size of the graphs increased. Therefore, in consideration of applying such a method to the biological domain, a minimum of 300 experiments should be included before attempting to build a small network.

II. Introduction

Since the human and mouse genome sequencing projects were completed, and advances in high throughput analytical technologies for assessing gene expression, there have been tremendous amounts of data produced on a vast array of topics. Because of this, new computational challenges are to make sense out of this overwhelming amount of experimental data. Of particular interest in this field though, is how to group genes with similar gene expression profiles together, and how to use this information to predict connections between genes to form a genetic network. One of the most interesting types of networks studied in this way are regulatory networks, which determine the cascade of events that must occur to initiate a metabolic process. Such networks would require directional connections between nodes, as well as many data points taken over a period of time. While this is a challenging, and as of yet unattainable task, it is currently the goal of many researchers to develop and apply these techniques within the context of genetics.

There are many different types of networks that can be created using gene expression data. The simplest of which is a correlation graph, where a Pearson or Spearman correlation coefficient is calculated for all pairs of genes on the array, and any coefficients above some threshold value are assumed to have an edge between them. From a biological standpoint though, this is too simplistic of a measure. It is only a pair-wise comparison between genes, which cannot account for multiple interactions among genes or non-linear relationships. As such, other methods have recently become popular in working with such data sets.

A popular algorithm for inferring underlying networks involves the application of Bayes theorem, and is known as a Bayesian Network. Such networks have been built

from biological data, and have been shown to be reliable and accurate [55, 56, 65, 96]. In their simplest sense, Bayesian networks provide a sophisticated measure of association between genes. They require both a prior and some new data, to calculate a posterior distribution. Algorithms for the construction of Bayesian networks are computationally intensive and rely on heuristic methods. Many different heuristics have been attempted, such as the K2 algorithm, Markov-Chain Monte Carlo methods and genetic algorithms.

One possible way to circumvent the computational complexity of Bayesian methods would be to identify groups of genes associated with a particular function from a larger set. Methods to do this, such as those used in Chapter 2 would be ideal. However, as we demonstrate, a large number of data points are required to learn the structure of a Bayesian network. Therefore the data set studied in Chapter 2 could not be utilized in this chapter. Instead, data was simulated, so that a known structure was available to use as a reference to test the accuracy of the structure learning methods.

III. Methods

Bayes Net Toolkit:

The Bayes Net Toolkit (BNT)[69] is an open-source toolbox for the development of directed graphical models for use in Matlab. The toolkit has methods that support many different types of conditional probability distributions, exact and approximate inference, parameter and structure learning, and static and dynamic models [69]. Of particular interest in this application are the functions supporting structure learning. The BNT has the ability for exhaustive search, which enumerates all possible graphs. There were three computers used in computing the graphical models in this section. The first two are sun4u Sun Fire 880 machines running sunOS 5.9 with four 900 Mhz processors and 8192 Megabytes of RAM. The third is a sun4u Sun Blade 100 (UltraSPARC-IIe) running sunOS 5.9 with one 502 Mhz processor and 512 Megabytes of RAM. The first two machines have Matlab Version 6.5.0.180913a Release 13, and the third has Matlab Version 7.0.1.24704 (R14) Service Pack 1 available. Given the aforementioned hardware platforms, it was not possible to enumerate through exact methods all possible graphs on a set with more than three nodes. Therefore, approximation methods were used. There are several available in the BNT: the K2 algorithm, Hill-climbing and the Metropolis-Hastings algorithm, which is a type of Markov-Chain Monte Carlo (MCMC) algorithm. All results obtained in this chapter were performed using the MCMC method, as it proved to give more accurate results (data not shown). The MCMC algorithm implemented in the BNT only allows a directed acyclic graph to be computed, and uses the acyclicity check described by Giudici et. al. [97].

Topology Simulation:

The topology of a graph describes the connection between its nodes along with the direction of the connection. The topology simulation entailed assessing the accuracy of the MCMC algorithm in learning the entire structure of the graph. A graph was first simulated from a power-law distribution [98, 99], as biological networks have been

shown to exhibit scale-free properties [100, 101]. The topology of scale-free networks is dominated by several highly connected hubs, or nodes with the largest degree [102]. The degree of each node in the graph was simulated by inverting a power law probability density:

$$\text{Number of Nodes} = \frac{(\beta * \mu_i)^{-1/\lambda}}{2} \qquad \textit{Formula 3.1}$$

where β controls the y-intercept of the log-log plot, μ_I is a random number from the continuous uniform distribution (Matlab unifrnd() function), and λ controls the number of edges in the graph [99]. After the degree of each node was calculated, edges were randomly added into the graph, until the desired degree of each node was reached. While most graphs consider the in-degree and out-degree to be separate entities, the degree here is the combination of the two. Each node takes only binary values indicating an on or active state and an off state. This approximation was considered so that as increasing the number of discrete levels possible for each node would have greatly increased the amount of data required to learn the graph [71].

Once the graph is simulated, parameters had to be created. This consisted of creating a conditional probability table for each node, which is done by sampling from the dirichlet distribution. Such a table dictates the probability that the child node has any possible value for all combinations of all possible values of the parent nodes [79]. The random number generator was seeded with three different integers for each of the three iterations: 1357901, 2930135 and 7012567. A tabular conditional probability table was used for each node, and from these tables, sample data was created. This was done using the BNT function sample_bnet.

Once data was simulated, the BNT's learn_struct_mcmc() function was used to search over the possible directed acyclic graphs. Each time this algorithm was run, 10,000 samples were taken from the possible search space, after 600 graphs were sampled (the burnin). The accept/reject ratio was returned from each of the sampled graphs. This was plotted for each simulation, a sample of which appears in *Figure 3.1*.

Each of the 10,000 sampled graphs were then scored using the Bayesian Information Criteria (BIC) scoring metric[66]. In the BNT, BIC is calculated as follows:

$$\log P(D \mid \text{theta_hat}) - 0.5*d*\log(N) \qquad \textit{Formula 3.2}$$

where D is the data, theta_hat is the maximum likelihood estimate of the parameters, d is the number of parameters, and N is the number of data cases. The BIC method has the advantage of not requiring a prior. The graph with the maximal score was kept, although it is recognized there may be multiple graphs with an equal score. Conditional probability tables were learned from the data for the highest scoring graph using the learn_params function, which attempts to find the maximum likelihood parameters for a fully observed model. Several different metrics were used to assess the quality of the

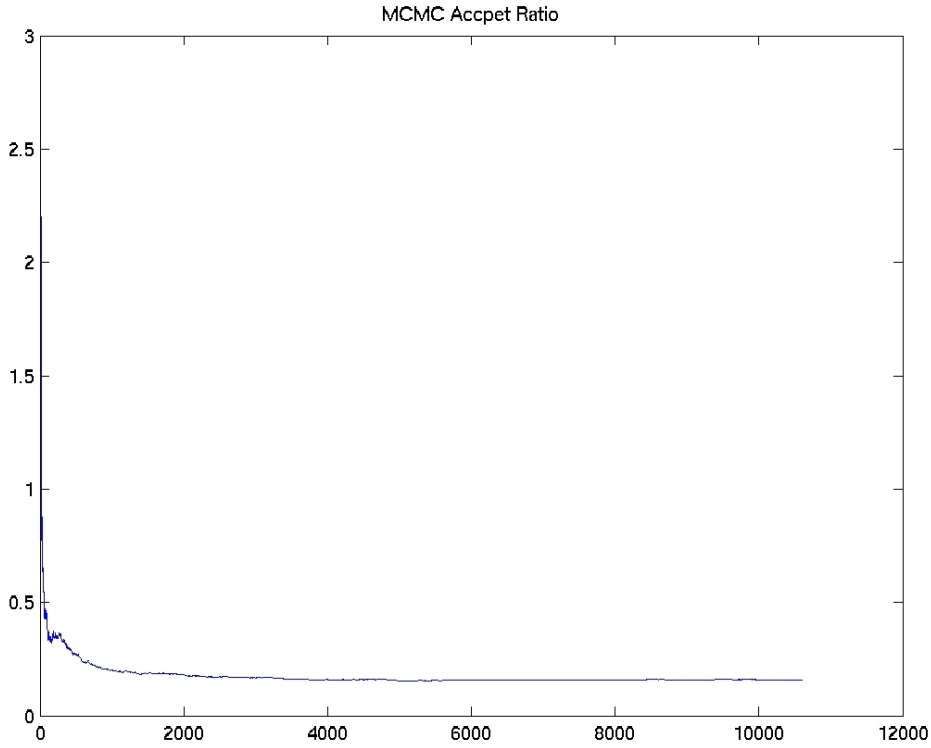


Figure 3.1: Sample plot of MCMC accept/reject ratio.

learned model versus the original model. The first four are just measures of the differences of edges in the learned graph versus the original graph. These are: true positives, true negatives, false positives and false negatives. From these four measures, there are several other metrics that can be calculated, as described in Schafer et. al. [103]. These are defined in *Table 3.1*. The final measurement made was to count the number of empty rows in the conditional probability tables for each node in the learned graph. This was considered to be another indicator of how much data was required to confidently learn the graph topology.

The topology simulation was setup to measure how much data was required to accurately learn the topology of a graph of size n. Four different graph sizes were used: 5, 8, 10 and 25 nodes. Graphs were created using a power-law distribution, as mentioned above, with beta set to 7.0 and lambda set to 7.5. Plots of the degree of nodes were created, to compare the distribution of the original graph to that of the highest scoring learned graph, one of which is shown in *Figure 3.2*. For each graph size, the simulation was run three times, using 2, 4, 6, 8, 10, 25, 50, 75, 100, 300, 500, 700, 900 and 1500 data points. The metrics described above were measured in each iteration, and the three measurements for each node size and number of data points were averaged.

Table 3. 1: Definition of metrics used for assessing accuracy of the topological structure of a learned graph.

Quantity	Definition
Number of true edges	$TP + FN = \eta_A E$
Number of zero-edges	$TN + FP = \eta_O E$
Significant edges	$TP + FP = S$
False positive rate	$E(FP/(\eta_O E)) = \alpha_1$
False negative rate	$E(FN/(\eta_A E)) = \alpha_{11}$
True negative rate (specificity)	$1- \alpha_1$
True positive rate (sensitivity, power)	$1- \alpha_{11}$
Positive predictive value	$PPV = E(TP/S\|S > 0)$
False discovery rate	$FDR = E(FP/S\|S > 0)$

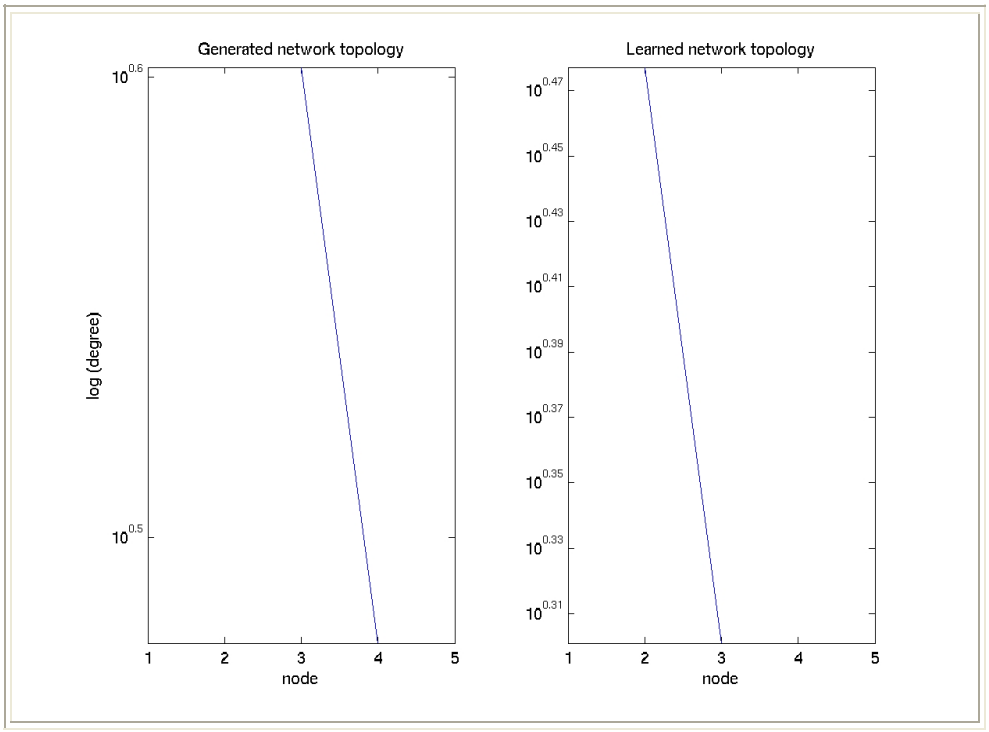


Figure 3.2: Sample plot of the node degree distribution of the original and learned graphs.

Features Simulation:

The features simulation entailed finding edges within high-scoring sampled graphs that were consistently present. In this simulation, a graph was generated, as described in the topology simulation section (above), as were conditional probability tables and sample data. 10,000 graphs were sampled with a burnin of 600 using the MCMC algorithm in the BNT. Each of the sampled graphs was scored using the BIC metric, and these were subsequently sorted. The scores from the returned samples were plotted, are shown in *Figure 3.3*. As with the topology simulation, there were 4 different sizes of graphs used: 5, 8, 10 and 25 nodes. There were seven different sizes of data points used: 100, 300, 500, 700, 900, 1500 and 5000. Each of the four graphs was generated using the aforementioned number of data points five times. The seeds for the random number generator are as follows: 1357901, 2930135, 7012567, 8225410 and 5552431.

Once the scores from the sampled graphs were sorted, those graphs with a score within 3 of the highest scoring graph were retained. A matrix was then created from these, where the cells represented the proportion of graphs within this sample that contain an edge. Since this was done five times, these were averaged to create a final "features" matrix. Any edges that were present in 60% of the graphs or more were retained, and those were used to create a graph. The features graphs, and original graphs were then visualized in ALIVE and compared.

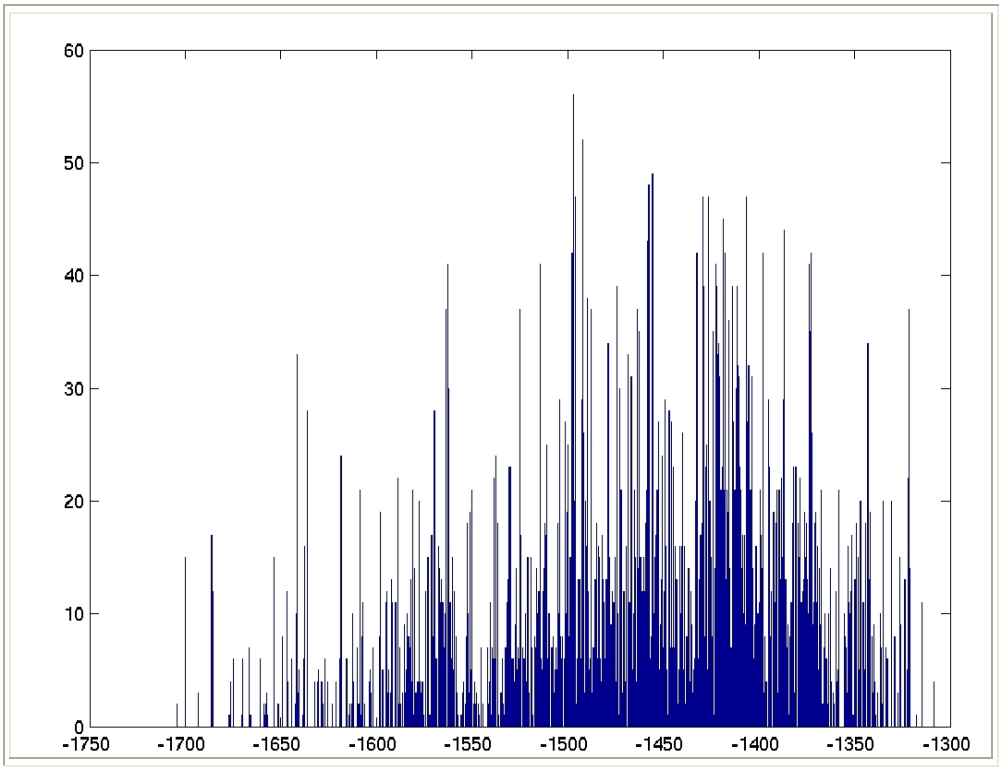


Figure 3.3: Example histogram of BIC scores from the features simulation. Scores were binned into groups of 3.

ALIVE:

Applet for Looking at Vertices and Edges (ALIVE) is an interactive graph visualization tool for multidimensional views of biological networks developed by Xinxia Peng and Adam Tebbe. It allows for the interactive isolation of small subgraphs with a large network in various ways, so as to allow users to focus on defined regions of interest. Additionally, it allows for multidimensional relationships between nodes to be identified, using node and edge coloring. A node or edge can be labeled with multiple values, so that attributes in common between nodes can be visually identified. ALIVE has several useful features for graph manipulation, but also allows the user to build the graph from a single node, or "walk" through the graph, by selecting the next node of interest to add to the graph. A screenshot of ALIVE is shown in *Figure 3.4*, The nodes are labeled with the gene names, and are colored according their classification as listed on BioCarta's website. The edges are labeled with the interaction classification also listed on BioCarta's website. They are colored according to their labels (although could be colored by a different label, again allowing for multidimensional data to be viewed in a complementary fashion). Also shown is the node information for BCAR1. This shows all of the labels for the node, in addition to all of its neighbors (which in this case is only PTK2).

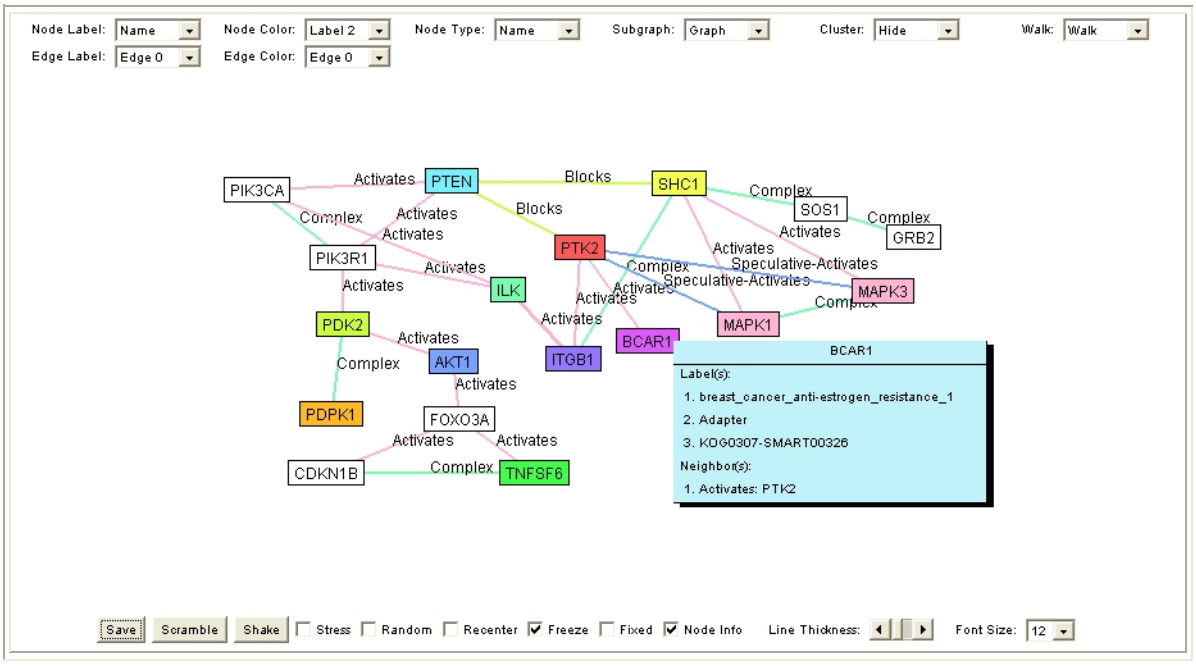


Figure 3. 4: Screenshot of ALIVE's interface.

IV. Results

Topology Simulation

The following four tables display the different metrics of accuracy of the Markov-Chain Monte Carlo algorithm in the Bayes Net Toolkit. *Table 3.2* contains the results from a graph containing 5 nodes, with the number of simulated data points ranging from 2 to 5000. *Tables 3.3, 3.4 and 3.5* display these same results for graphs with 8, 10 and 25 nodes respectively. *Figures 3.5, 3.6, 3.7 and 3.8* contain plots of these different results, for a visual interpretation of these results.

Features Simulations

Table 3.7 displays the results of the 5 runs for each level of data points in the features simulation of a graph with 5 nodes. Increasing the number of data points used did not have a positive effect on the highest scoring graph, or number of graphs that were within 3 of the highest scoring graph. The seed for each of the 5 runs was different, although with each number of data points in the simulated data, the same seeds were used to initialize the random number generator. In *Table 3.6*, the average features matrix is created by averaging the number of graphs with any given edge. *Table 3.6* was created from the graphs sampled with 5000 points of simulated data. The number of graphs used to create the matrix was 18,233. *Figure 3.9* displays the structure of the original graph on the right (green), and the graph with edges present in 60% or more of the sampled graphs on the left (red).

Table 3.9 displays the results of the 5 runs for each level of data points in the features simulation of a graph with 8 nodes. Increasing the number of data points beyond 300 did not seem to have a positive effect on the highest scoring graph, or the number of graphs that were within 3 of the highest scoring graph. The seed for each of the 5 runs was different, although with each number of data points in the simulated data, the same seeds were used to initialize the random number generator. In *Table 3.8*, the average features matrix is created by averaging the number of graphs with any given edge. *Table 3.8* was created from the graphs sampled with 5000 data points of simulated data. The number of graphs used to create the matrix was 177. *Figure 3.10* displays the original graph on the right (green), and the graph with edges appearing in 60% or more of the sampled graphs on the left (red).

Table 3.11 displays the results of the 5 runs for each level of data points in the features simulation of a graph with 10 nodes. Increasing the number of data points beyond 300 did not seem to have a positive effect on the highest scoring graph, or the number of graphs that were within 3 of the highest scoring graph. The seed for each of the 5 iterations was different, although with each number of data points in the simulated data, the same seeds were used to initialize the random number generator. In *Table 3.10*, the average features matrix is created by averaging the number of graphs with any given edge. *Table 3.10* was created from the graphs sampled with 5000 data points of simulated data. The number of graphs used to create the matrix was 207. *Figure 3.11*

Table 3.2: Metrics of accuracy for a graph with 5 nodes.

5 Nodes	2	4	6	8	10	25	50	75	100	300	500	700	900	1500	5000
True Positives	0.667	1.333	1.000	2.333	2.000	2.000	1.000	1.667	1.667	3.000	3.667	3.333	3.333	3.667	4.333
True Negatives	17.000	16.667	16.667	17.000	16.333	16.667	15.333	15.667	16.000	16.000	16.000	15.667	15.667	15.667	16.000
False Positives	1.000	1.333	1.333	1.000	1.667	1.333	2.667	2.333	2.000	2.000	2.000	2.333	2.333	2.333	2.000
False Negatives	6.333	5.667	6.000	4.667	5.000	5.000	6.000	5.333	5.333	4.000	3.333	3.667	3.667	3.333	2.667
Empty CPT Rows	3.000	1.667	2.000	1.000	0.667	0.333	0.333	0.333	0.333	0.333	0.000	0.000	0.333	0.000	0.000
True Edges	7.000	7.000	7.000	7.000	7.000	7.000	7.000	7.000	7.000	7.000	7.000	7.000	7.000	7.000	7.000
Zero Edges	18.000	18.000	18.000	18.000	18.000	18.000	18.000	18.000	18.000	18.000	18.000	18.000	18.000	18.000	18.000
Significant Edges	1.667	2.667	2.333	3.333	3.667	3.333	3.667	4.000	3.667	5.000	5.667	5.667	5.667	6.000	6.333
False Positive Rate	0.056	0.074	0.074	0.056	0.093	0.074	0.148	0.130	0.111	0.111	0.111	0.130	0.130	0.130	0.111
False Negative Rate	0.905	0.810	0.857	0.667	0.714	0.714	0.857	0.762	0.762	0.571	0.476	0.524	0.524	0.476	0.381
True Positive Rate	0.095	0.190	0.143	0.333	0.286	0.286	0.143	0.238	0.238	0.429	0.524	0.476	0.476	0.524	0.619
True Negative Rate	0.944	0.926	0.926	0.944	0.907	0.926	0.852	0.870	0.889	0.889	0.889	0.870	0.870	0.870	0.889
Positive Predictive Value	0.400	0.500	0.429	0.700	0.545	0.600	0.273	0.417	0.455	0.600	0.647	0.588	0.588	0.611	0.684
False Discovery Rate	0.600	0.500	0.571	0.300	0.455	0.400	0.727	0.583	0.545	0.400	0.353	0.412	0.412	0.389	0.316

Table 3.3: Metrics of accuracy for a graph with 8 nodes.

8 Nodes	2	4	6	8	10	25	50	75	100	300	500	700	900	1500	5000
True Positives	1.667	1.333	2.000	3.667	2.333	3.667	4.000	4.000	3.667	6.333	7.333	7.667	6.333	8.000	8.000
True Negatives	45.000	44.000	45.333	46.000	45.667	46.000	45.000	45.000	44.000	46.000	46.333	46.667	45.333	46.667	45.667
False Positives	5.000	6.000	4.667	4.000	4.333	4.000	5.000	5.000	6.000	4.000	3.667	3.333	4.667	3.333	4.333
False Negatives	12.333	12.667	12.000	10.333	11.667	10.333	10.000	10.000	10.333	7.667	6.667	6.333	7.667	6.000	6.000
Empty CPT Rows	6.667	5.333	6.000	3.667	2.000	1.333	0.333	1.333	1.000	1.000	1.333	2.000	2.000	2.333	1.333
True Edges	14.000	14.000	14.000	14.000	14.000	14.000	14.000	14.000	14.000	14.000	14.000	14.000	14.000	14.000	14.000
Zero Edges	50.000	50.000	50.000	50.000	50.000	50.000	50.000	50.000	50.000	50.000	50.000	50.000	50.000	50.000	50.000
Significant Edges	6.667	7.333	6.667	7.667	6.667	7.667	9.000	9.000	9.667	10.333	11.000	11.000	11.000	11.333	12.333
False Positive Rate	0.100	0.120	0.093	0.080	0.087	0.080	0.100	0.100	0.120	0.080	0.073	0.067	0.093	0.067	0.087
False Negative Rate	0.881	0.905	0.857	0.738	0.833	0.738	0.714	0.714	0.738	0.548	0.476	0.452	0.548	0.429	0.429
True Positive Rate	0.119	0.095	0.143	0.262	0.167	0.262	0.286	0.286	0.262	0.452	0.524	0.548	0.452	0.571	0.571
True Negative Rate	0.900	0.880	0.907	0.920	0.913	0.920	0.900	0.900	0.880	0.920	0.927	0.933	0.907	0.933	0.913
Positive Predictive Value	0.250	0.182	0.300	0.478	0.350	0.478	0.444	0.444	0.379	0.613	0.667	0.697	0.576	0.706	0.649
False Discovery Rate	0.750	0.818	0.700	0.522	0.650	0.522	0.556	0.556	0.621	0.387	0.333	0.303	0.424	0.294	0.351

Table 3.4: Metrics of accuracy for a graph with 10 nodes.

10 Nodes	2	4	6	8	10	25	50	75	100	300	500	700	900	1500	5000
True Positives	3.000	3.000	1.333	3.000	4.667	2.667	4.000	3.333	3.667	7.333	4.333	7.000	9.333	8.000	10.000
True Negatives	71.000	71.667	70.333	71.667	74.000	73.667	76.333	75.333	75.000	78.667	74.000	78.333	79.667	78.000	79.667
False Positives	12.000	11.333	12.667	11.333	9.000	9.333	6.667	7.667	8.000	4.333	9.000	4.667	3.333	5.000	3.333
False Negatives	14.000	14.000	15.667	14.000	12.333	14.333	13.000	13.667	13.333	9.667	12.667	10.000	7.667	9.000	7.000
Empty CPT Rows	20.000	21.000	13.000	14.000	8.000	4.333	2.667	4.667	4.000	4.667	3.000	3.333	0.667	2.333	3.000
True Edges	17.000	17.000	17.000	17.000	17.000	17.000	17.000	17.000	17.000	17.000	17.000	17.000	17.000	17.000	17.000
Zero Edges	83.000	83.000	83.000	83.000	83.000	83.000	83.000	83.000	83.000	83.000	83.000	83.000	83.000	83.000	83.000
Significant Edges	15.000	14.333	14.000	14.333	13.667	12.000	10.667	11.000	11.667	11.667	13.333	11.667	12.667	13.000	13.333
False Positive Rate	0.145	0.137	0.153	0.137	0.108	0.112	0.080	0.092	0.096	0.052	0.108	0.056	0.040	0.060	0.040
False Negative Rate	0.824	0.824	0.922	0.824	0.725	0.843	0.765	0.804	0.784	0.569	0.745	0.588	0.451	0.529	0.412
True Positive Rate	0.176	0.176	0.078	0.176	0.275	0.157	0.235	0.196	0.216	0.431	0.255	0.412	0.549	0.471	0.588
True Negative Rate	0.855	0.863	0.847	0.863	0.892	0.888	0.920	0.908	0.904	0.948	0.892	0.944	0.960	0.940	0.960
Positive Predictive Value	0.200	0.209	0.095	0.209	0.341	0.222	0.375	0.303	0.314	0.629	0.325	0.600	0.737	0.615	0.750
False Discovery Rate	0.800	0.791	0.905	0.791	0.659	0.778	0.625	0.697	0.686	0.371	0.675	0.400	0.263	0.385	0.250

Table 3.5: Metrics of accuracy for a graph with 25 nodes.

25 Nodes	2	4	6	8	10	25	50	75	100	300	500	700	900	1500	5000
True Positives	8.667	12.000	10.000	9.333	11.333	11.000	9.333	13.000	13.667	24.000	27.667	31.667	31.333	32.000	36.333
True Negatives	446.667	450.333	452.333	457.333	461.667	508.333	525.000	534.000	534.667	548.667	550.333	553.000	556.667	552.667	547.333
False Positives	127.333	123.667	121.667	116.667	112.333	65.667	49.000	40.000	39.333	25.333	23.667	21.000	17.333	21.333	26.667
False Negatives	42.333	39.000	41.000	41.667	39.667	40.000	41.667	38.000	37.333	27.000	23.333	19.333	19.667	19.000	14.667
Empty CPT Rows	6764.333	5386.000	4716.667	4095.333	4892.667	655.333	113.000	58.667	54.000	17.333	22.000	18.000	13.000	12.667	25.333
True Edges	51.000	51.000	51.000	51.000	51.000	51.000	51.000	51.000	51.000	51.000	51.000	51.000	51.000	51.000	51.000
Zero Edges	574.000	574.000	574.000	574.000	574.000	574.000	574.000	574.000	574.000	574.000	574.000	574.000	574.000	574.000	574.000
Significant Edges	136.000	135.667	131.667	126.000	123.667	76.667	58.333	53.000	53.000	49.333	51.333	52.667	48.667	53.333	63.000
False Positive Rate	0.222	0.215	0.212	0.203	0.196	0.114	0.085	0.070	0.069	0.044	0.041	0.037	0.030	0.037	0.046
False Negative Rate	0.830	0.765	0.804	0.817	0.778	0.784	0.817	0.745	0.732	0.529	0.458	0.379	0.386	0.373	0.288
True Positive Rate	0.170	0.235	0.196	0.183	0.222	0.216	0.183	0.255	0.268	0.471	0.542	0.621	0.614	0.627	0.712
True Negative Rate	0.778	0.785	0.788	0.797	0.804	0.886	0.915	0.930	0.931	0.956	0.959	0.963	0.970	0.963	0.954
Positive Predictive Value	0.064	0.088	0.076	0.074	0.092	0.143	0.160	0.245	0.258	0.486	0.539	0.601	0.644	0.600	0.577
False Discovery Rate	0.936	0.912	0.924	0.926	0.908	0.857	0.840	0.755	0.742	0.514	0.461	0.399	0.356	0.400	0.423

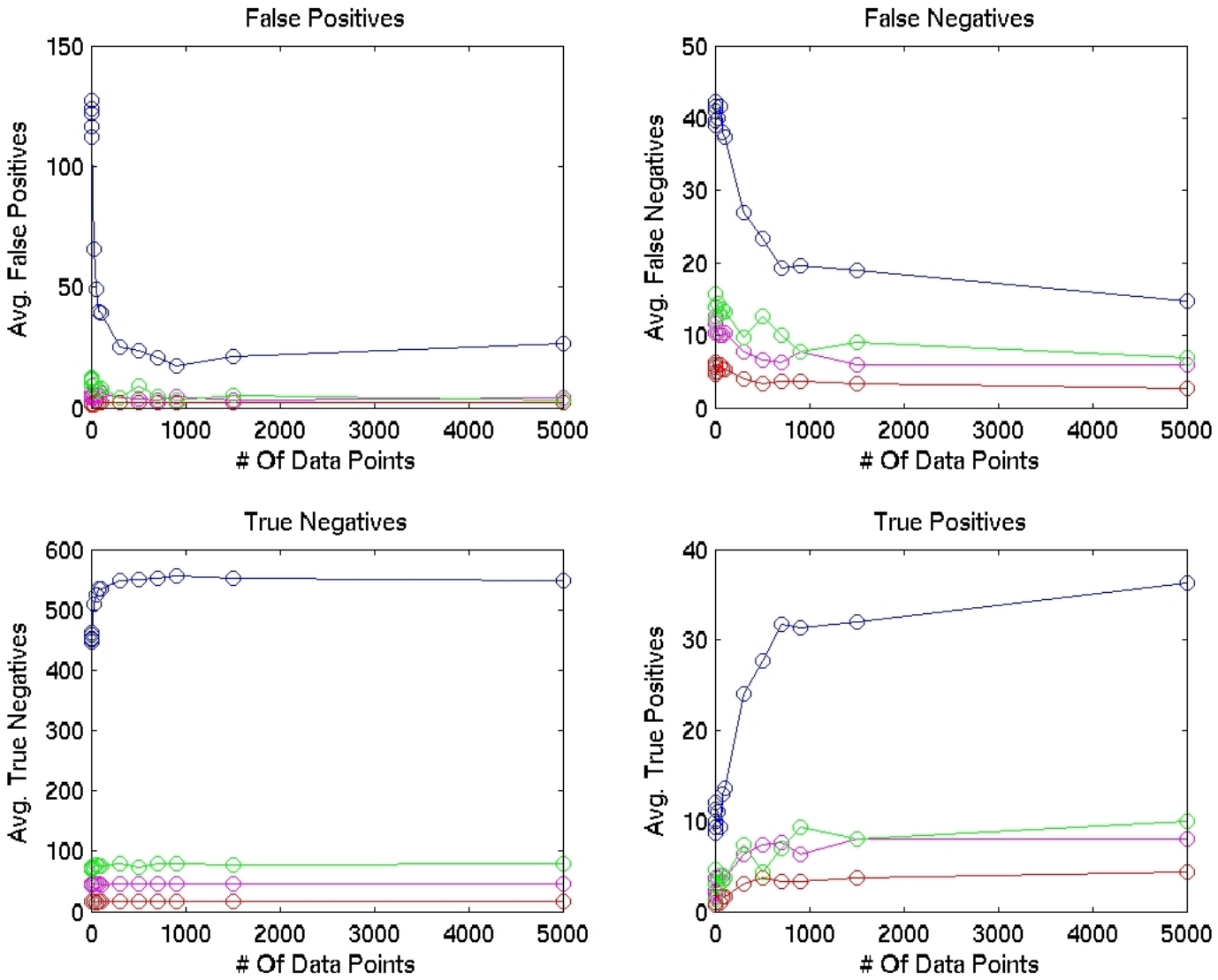


Figure 3.5: Average True Positives, True Negatives, False Positives, False Negatives of highest scoring graphs with a given number of nodes. Blue line: 25 nodes, Green line: 10 nodes, Purple line: 8 nodes, Red line: 5 nodes.

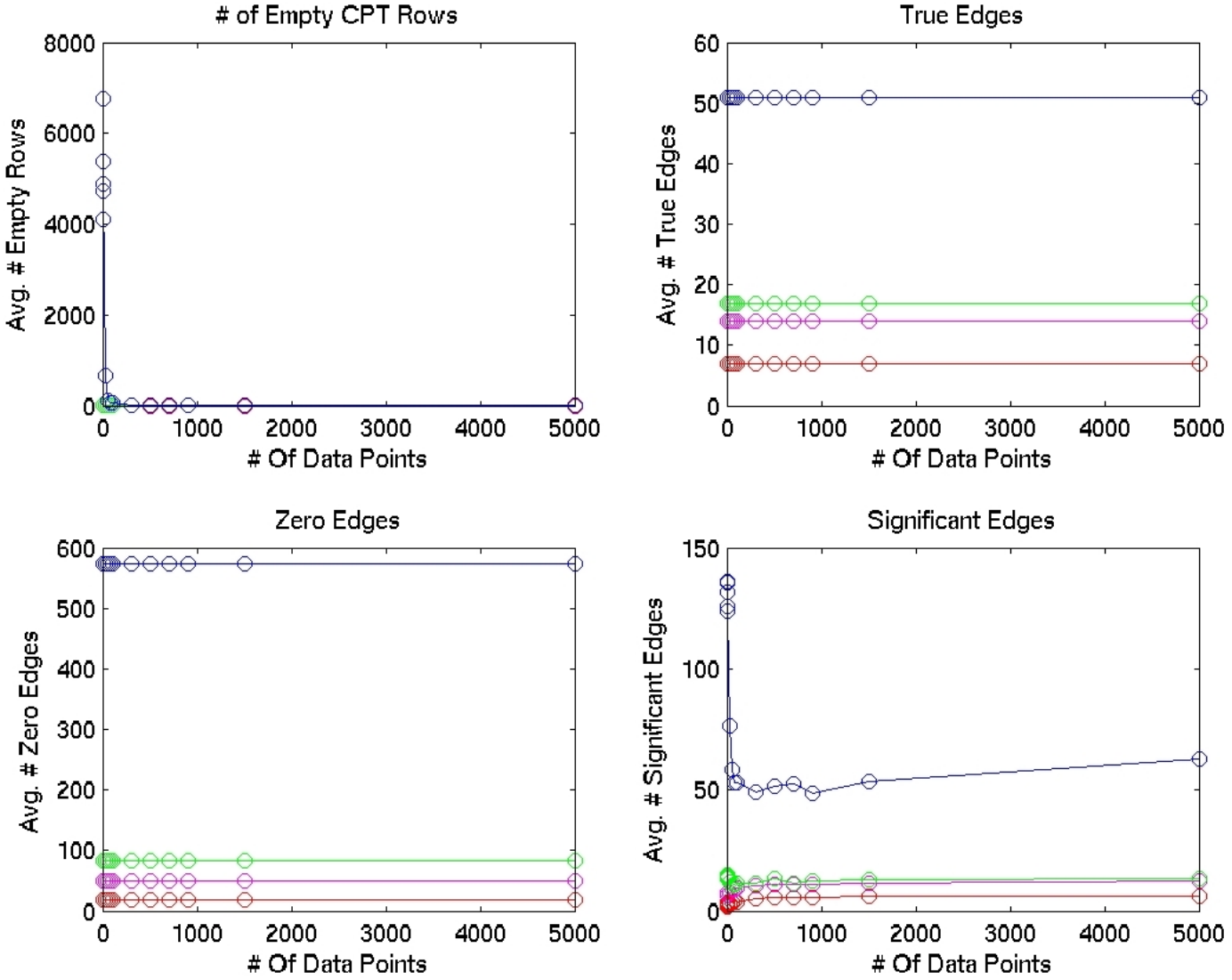


Figure 3.6: Average Empty Rows, True Edges, Zero Edges and Significant Edges of highest scoring graphs with a given number of nodes. Blue line: 25 nodes, Green line: 10 nodes, Purple line: 8 nodes, Red line: 5 nodes.

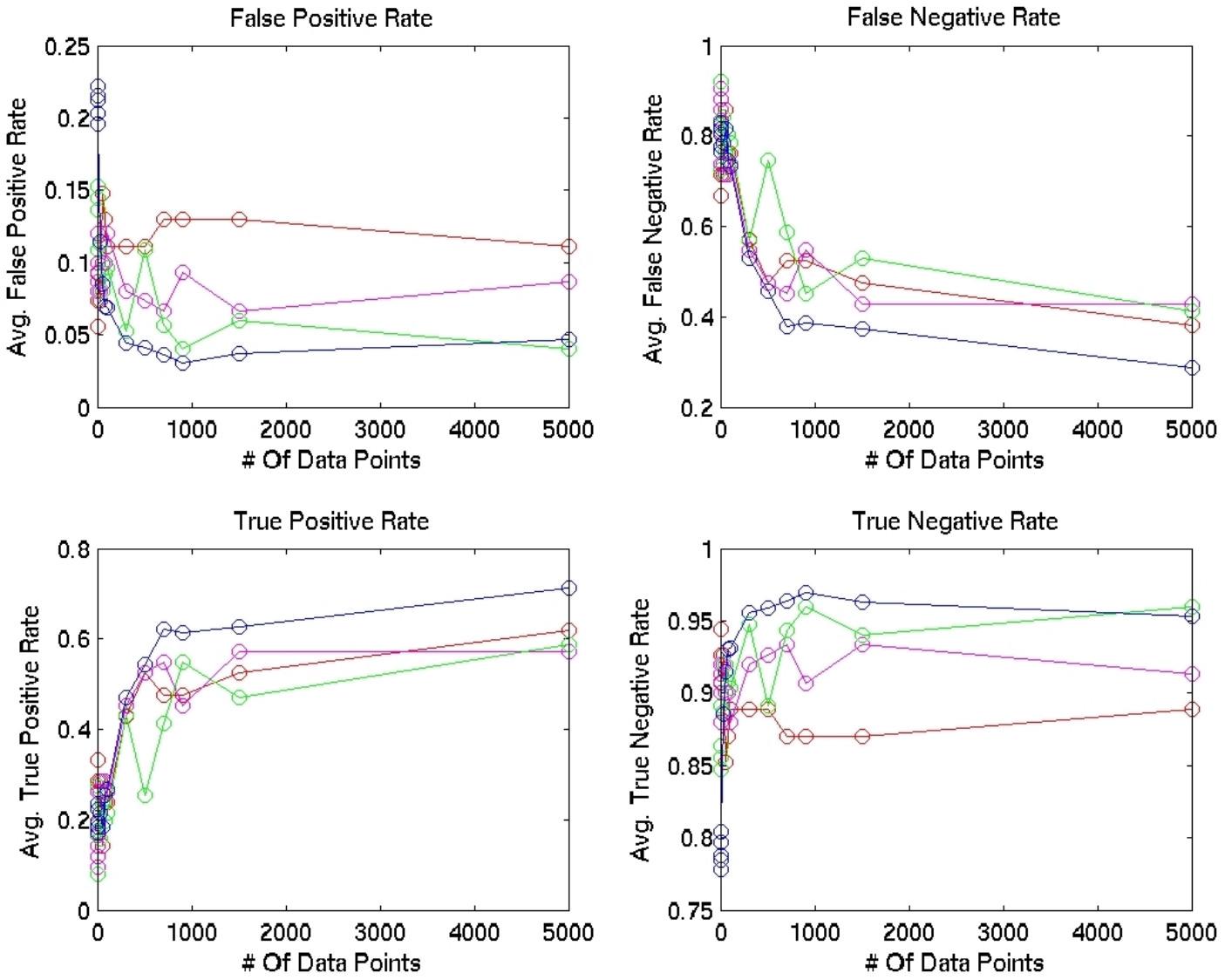


Figure 3.7: Average False Positive Rate, False Negative Rate, True Positive Rate and True Negative Rate of highest scoring graphs with a given number of nodes. Blue line: 25 nodes, Green line: 10 nodes, Purple line: 8 nodes, Red line: 5 nodes.

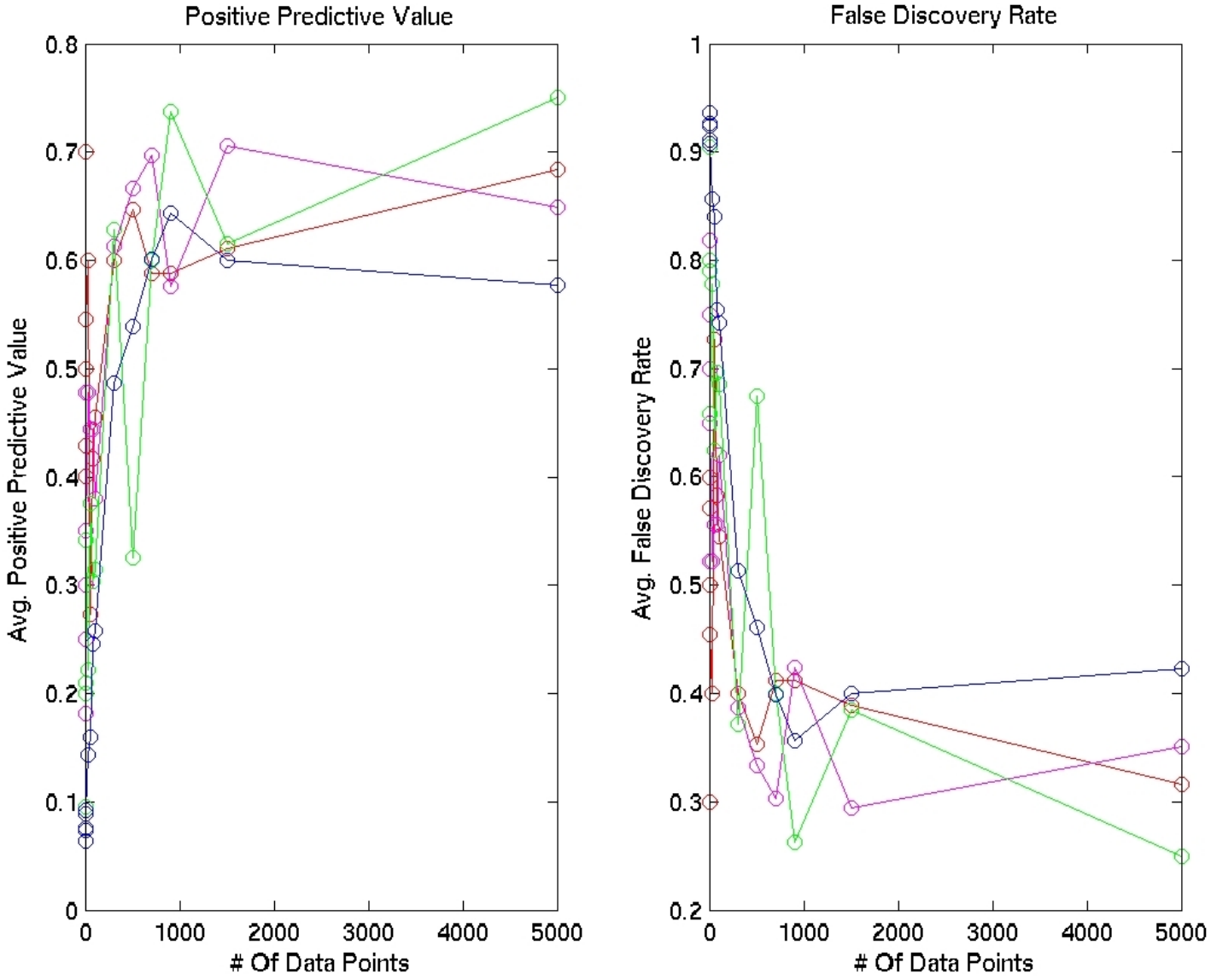


Figure 3.8: Average Positive Predictive Value and False Discovery Rate of highest scoring graphs with a given number of nodes. Blue line: 25 nodes, Green line: 10 nodes, Purple line: 8 nodes, Red line: 5 nodes.

Table 3.6: Average features matrix of graph with 5 nodes from 5000 data points.

0.000000	0.652169	0.467065	0.054133	0.380683
0.158010	0.000000	0.004442	0.831405	0.347173
0.532935	0.006691	0.000000	0.032743	0.088850
0.036637	0.168595	0.000000	0.000000	0.000000
0.619317	0.652827	0.454067	0.051116	0.000000

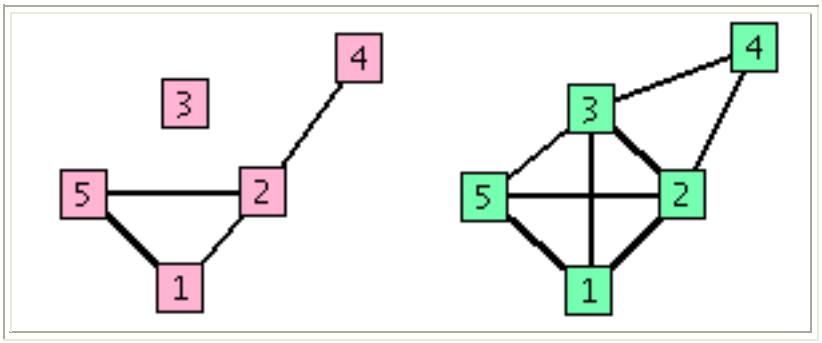


Figure 3.9: Graph from features matrix with 5 nodes from 5000 data points. Edges were included where at least 60% of sampled graphs had an edge.

Table 3.7: Results of features simulation of graphs with 5 nodes. The highest scoring graph is shown, as well as the number of graphs within 3 of the highest scoring graph, and the score the graph with the lowest score within 3 of the highest scoring graph.

Data Points	Highest Score	# of Graphs	Lowest score
100	-230.569480	3828	-233.333649
	-230.569480	3464	-233.333649
	-230.569480	3557	-233.333649
	-230.569480	3628	-233.333649
	-230.569480	3756	-233.333649
300	-230.569480	3823	-233.333649
	-230.569480	3464	-233.333649
	-230.569480	3557	-233.333649
	-230.569480	3628	-233.333649
	-230.569480	3756	-233.333649
500	-230.569480	3823	-233.333649
	-230.569480	3464	-233.333649
	-230.569480	3557	-233.333649
	-230.569480	3628	-233.333649
	-230.569480	3756	-233.333649
700	-230.569480	3823	-233.333649
	-230.569480	3464	-233.333649
	-230.569480	3557	-233.333649
	-230.569480	3628	-233.333649
	-230.569480	3756	-233.333649
900	-230.569480	3823	-233.333649
	-230.569480	3464	-233.333649
	-230.569480	3557	-233.333649
	-230.569480	3628	-233.333649
	-230.569480	3756	-233.333649
1500	-230.569480	3823	-233.333649
	-230.569480	3464	-233.333649
	-230.569480	3557	-233.333649
	-230.569480	3628	-233.333649
	-230.569480	3756	-233.333649
5000	-230.569480	3823	-233.333649
	-230.569480	3464	-233.333649
	-230.569480	3557	-233.333649
	-230.569480	3628	-233.333649
	-230.569480	3756	-233.333649

Table 3.8: Average features matrix of graph with 8 nodes from 5000 data points.

0.000000	0.000000	0.932203	0.079096	0.000000	0.468927	0.045198	0.502825
0.000000	0.000000	0.000000	0.022599	0.000000	0.000000	0.000000	0.000000
0.067797	0.000000	0.000000	0.101695	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.016949	0.000000	0.011299	0.000000
0.028249	1.000000	0.548023	0.079096	0.000000	0.000000	0.000000	0.231638
0.531073	0.000000	0.000000	0.073446	0.000000	0.000000	0.576271	0.000000
0.000000	1.000000	0.000000	0.039548	0.000000	0.327684	0.000000	0.000000
0.497175	1.000000	0.000000	0.056497	0.706215	0.000000	0.045198	0.000000

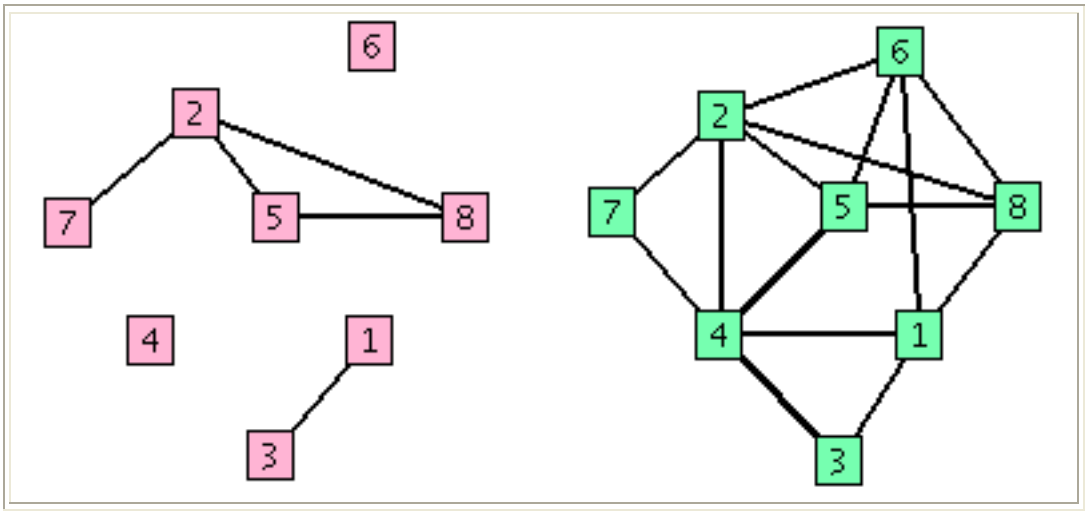


Figure 3.10: Graph from features matrix with 8 nodes from 5000 data points. Edges were included where at least 60% of sampled graphs had an edge.

Table 3.9: Results of features simulation of graphs with 8 nodes. The highest scoring graph is shown, as well as the number of graphs within 3 of the highest scoring graph, and the score the graph with the lowest score within 3 of the highest scoring graph.

Data Points	Highest Score	# of Graphs	Lowest score
100	-329.228589	57	-332.174279
	-328.403125	38	-331.050571
	-329.871694	60	-332.805493
	-329.505107	25	-332.174279
	-328.403125	35	-330.973676
300	-328.403125	14	-330.705710
	-328.403125	38	-331.050571
	-329.871694	60	-332.805493
	-329.505107	25	-332.174279
	-328.403125	35	-330.973676
500	-328.403125	14	-330.705710
	-328.403125	38	-331.050571
	-329.871694	60	-332.805493
	-329.505107	25	-332.174279
	-328.403125	35	-330.973676
700	-328.403125	14	-330.705710
	-328.403125	38	-331.050571
	-329.871694	60	-332.805493
	-329.505107	25	-332.174279
	-328.403125	35	-330.973676
900	-328.403125	14	-330.705710
	-328.403125	38	-331.050571
	-329.871694	60	-332.805493
	-329.505107	25	-332.174279
	-328.403125	35	-330.973676
1500	-328.403125	14	-330.705710
	-328.403125	38	-331.050571
	-329.871694	60	-332.805493
	-329.505107	25	-332.174279
	-328.403125	35	-330.973676
5000	-328.403125	14	-330.705710
	-328.403125	38	-331.050571
	-329.871694	60	-332.805493
	-329.505107	25	-332.174279
	-328.403125	35	-330.973676

Table 3.10: Average features matrix of graph with 10 nodes from 5000 data points.

0.000000	0.115942	0.275362	0.898551	0.000000	0.333333	0.000000	0.024155	0.000000	0.028986
0.149758	0.000000	0.135266	0.062802	0.178744	0.000000	0.000000	0.000000	0.294686	0.483092
0.724638	0.004831	0.000000	0.000000	0.164251	0.850242	0.043478	0.159420	0.589372	0.159420
0.101449	0.000000	0.000000	0.000000	0.096618	0.028986	0.000000	0.922705	0.000000	0.000000
0.000000	0.057971	0.096618	0.173913	0.000000	0.000000	0.101449	0.000000	0.053140	0.193237
0.565217	0.000000	0.149758	0.000000	0.164251	0.000000	0.000000	0.178744	0.000000	0.000000
0.000000	0.000000	0.004831	0.000000	0.004831	0.000000	0.000000	0.000000	0.000000	0.009662
0.000000	0.000000	0.000000	0.077295	0.101449	0.000000	0.062802	0.000000	0.000000	0.000000
0.000000	0.024155	0.280193	0.000000	0.019324	0.000000	0.000000	0.033816	0.000000	0.164251
0.719807	0.516908	0.420290	0.000000	0.086957	0.000000	0.019324	0.000000	0.294686	0.000000

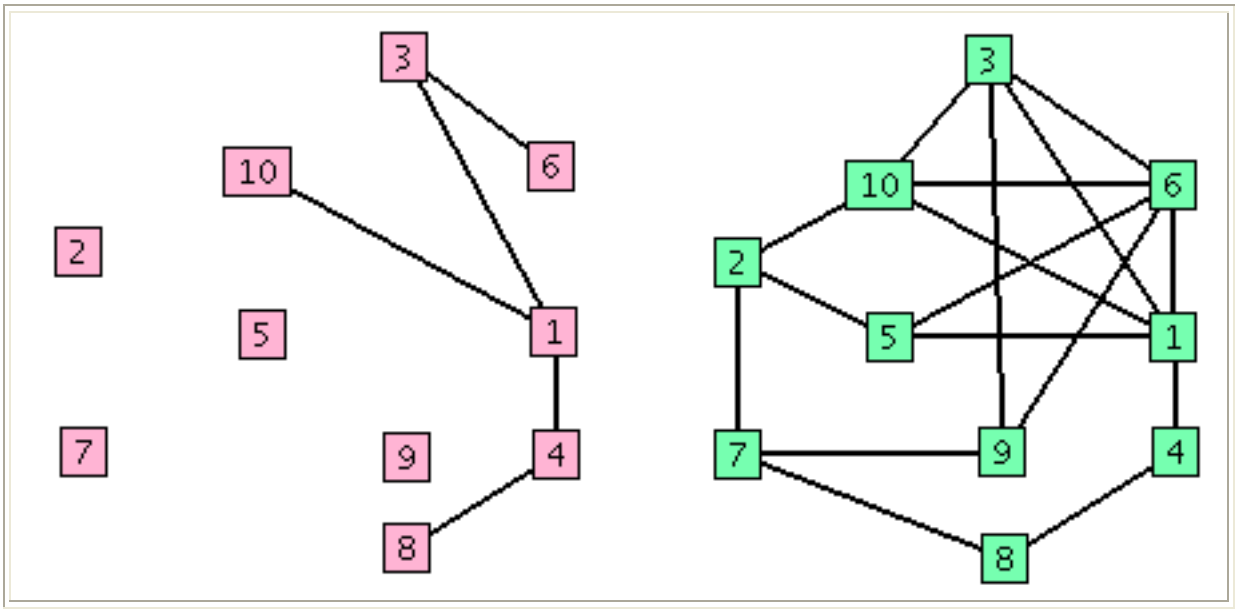


Figure 3.11: Graph from features matrix with 10 nodes from 5000 data points. Edges were included where at least 60% of sampled graphs had an edge.

Table 3.11: Results of features simulation of graphs with 10 nodes. The highest scoring graph is shown, as well as the number of graphs within 3 of the highest scoring graph, and the score the graph with the lowest score within 3 of the highest scoring graph.

Data Points	Highest Score	# of Graphs	Lowest score
100	-397.901706	15	-400.477969
	-402.539461	35	-405.244542
	-402.780554	35	-405.083139
	-401.840493	56	-404.620754
	-402.049110	11	-404.665614
300	-403.001455	65	-405.304041
	-402.539461	35	-405.244542
	-402.780554	35	-405.083139
	-401.840493	56	-404.620754
	-402.049110	11	-404.665614
500	-403.001455	65	-405.304041
	-402.539461	35	-405.244542
	-402.780554	35	-405.083139
	-401.840493	56	-404.620754
	-402.049110	11	-404.665614
700	-403.001455	65	-405.304041
	-402.539461	35	-405.244542
	-402.780554	35	-405.083139
	-401.840493	56	-404.620754
	-402.049110	11	-404.665614
900	-403.001455	65	-405.304041
	-402.539461	35	-405.244542
	-402.780554	35	-405.083139
	-401.840493	56	-404.620754
	-402.049110	11	-404.665614
1500	-403.001455	65	-405.304041
	-402.539461	35	-405.244542
	-402.780554	35	-405.083139
	-401.840493	56	-404.620754
	-402.049110	11	-404.665614
5000	-403.001455	65	-405.304041
	-402.539461	35	-405.244542
	-402.780554	35	-405.083139
	-401.840493	56	-404.620754
	-402.049110	11	-404.665614

displays the original graph on the right (green), and the graph with edges appearing in 60% or more of the sampled graphs on the left (red).

Table 3.12 displays the results of the 5 runs for each level of data points in the features simulation of a graph with 25 nodes. Increasing the number of data points beyond 300 did not seem to have a positive effect on the highest scoring graph, or the number of graphs that were within 3 of the highest scoring graph. The seed for each of the 5 runs was different, although with each number of data points in the simulated data, the same seeds were used to initialize the random number generator. In *Table 3.13*, the average features matrix is created by averaging the number of graphs with any given edge. *Table 3.13* was created from the graphs sampled with 5000 data points of simulated data. The number of graphs used to create the matrix was 36. *Figure 3.13* displays the original graph on the right (green), and the graph with edges appearing in 60% or more of the sampled graphs on the left (red). This is the only graph presented thus far in the features simulation where there are false positive edges present.

Shown in *Figure 3.12* is the average score of the highest scoring graph in the 5 runs for each level of the four different sized graphs, with the range of data points. It is obvious that increasing the number of data points above 100, or in one case, 300 does not improve the average highest score. *Figure 3.14* displays bar plots of the four metrics used to assess the accuracy of the graphs with edges present in 60% or more of the sampled graphs, made from 5000 data points of simulated data. The graph with 25 nodes was the only one that had false positives present. False negatives increased with node size. True negatives increased with node size as well, although true positives did not increase as much as would be expected.

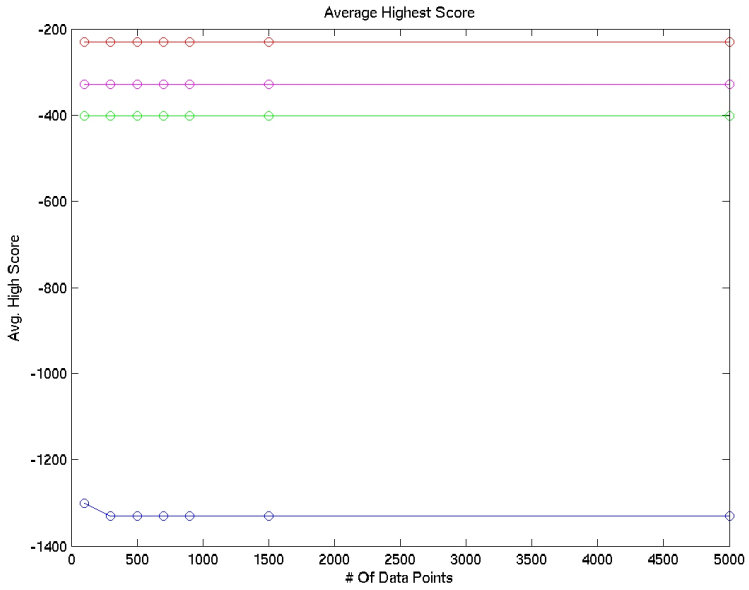


Figure 3.12: Average highest score of graphs in features simulation. Blue line: 25 nodes, Green line: 10 nodes, Purple line: 8 nodes, Red line: 5 nodes.

Table 3.12: Results of features simulation of graphs with 25 nodes. The highest scoring graph is shown, as well as the number of graphs within 3 of the highest scoring graph, and the score the graph with the lowest score within 3 of the highest scoring graph.

Data Points	Highest Score	# of Graphs	Lowest score
100	-1289.250427	29	-1291.752328
	-1326.141029	2	-1326.924291
	-1317.957726	3	-1317.957726
	-1264.591519	1	-1264.591519
	-1308.418787	3	-1308.418787
300	-1434.744052	22	-1437.035017
	-1326.141029	2	-1326.924291
	-1317.957726	3	-1317.957726
	-1264.591519	1	-1264.591519
	-1308.418787	3	-1308.418787
500	-1434.744052	22	-1437.035017
	-1326.141029	2	-1326.924291
	-1317.957726	3	-1317.957726
	-1264.591519	1	-1264.591519
	-1308.418787	3	-1308.418787
700	-1434.744052	22	-1437.035017
	-1326.141029	2	-1326.924291
	-1317.957726	3	-1317.957726
	-1264.591519	1	-1264.591519
	-1308.418787	3	-1308.418787
900	-1434.744052	22	-1437.035017
	-1326.141029	2	-1326.924291
	-1317.957726	3	-1317.957726
	-1264.591519	1	-1264.591519
	-1308.418787	3	-1308.418787
1500	-1434.744052	22	-1437.035017
	-1326.141029	2	-1326.924291
	-1317.957726	3	-1317.957726
	-1264.591519	1	-1264.591519
	-1308.418787	3	-1308.418787
5000	-1434.744052	22	-1437.035017
	-1326.141029	2	-1326.924291
	-1317.957726	3	-1317.957726
	-1264.591519	1	-1264.591519
	-1308.418787	3	-1308.418787

Table 3.13: Average features matrix of graph with 25 nodes from 5000 data points.

0.000	0.111	0.111	0.111	0.111	0.639	0.000	0.889	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.083	0.056	0.000	0.000	0.000	0.000	0.000
0.000	0.000	0.111	0.000	0.000	0.111	0.000	0.000	0.000	0.111	0.000	0.056	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.111	0.000	0.000	0.167
0.000	0.639	0.000	0.000	0.000	0.000	0.000	0.000	0.639	0.000	0.000	0.111	0.028	0.000	0.083	0.000	0.000	0.361	0.000	0.000	0.000	0.000	0.000	0.000	0.250
0.056	0.000	0.000	0.000	0.639	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.194	0.000	0.111	0.000	0.000
0.000	0.000	0.000	0.056	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.083	0.250	0.000	0.000	0.000	0.000	0.639	0.083	0.083	0.000	0.000	0.000	0.750	0.000	0.000	0.083	0.556	0.000	0.000	0.167	0.083	0.000	0.000	0.639	0.000
0.000	0.000	0.000	0.000	0.694	0.000	0.000	0.000	0.000	0.694	0.000	0.000	0.750	0.000	0.000	0.194	0.000	0.000	0.000	0.000	0.000	0.167	0.750	0.806	0.000
0.111	0.000	0.000	0.194	0.194	0.556	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.806	0.111	0.000	0.000	0.083	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.194	0.000	0.000	0.611	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.361	0.000	0.000	0.000	0.111	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.194	0.000	0.111	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.111	0.000	0.667	0.000
0.111	0.000	0.306	0.000	0.000	0.639	0.111	0.194	0.639	0.111	0.000	0.111	0.000	0.639	0.833	0.111	0.667	0.000	0.000	0.083	0.000	0.722	0.722	0.111	0.111
0.806	0.639	0.111	0.000	0.694	0.083	0.083	0.722	0.111	0.000	0.750	0.000	0.194	0.000	0.556	0.722	0.083	0.000	0.639	0.000	0.111	0.750	0.639	0.111	0.167
0.000	0.000	0.000	0.639	0.000	0.250	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.000	0.250	0.000	0.000	0.000	0.000	0.000	0.056	0.000	0.111	0.000	0.139	0.000	0.000	0.750	0.000	0.000	0.000	0.000	0.000	0.000	0.278	0.000	0.000	0.000
0.639	0.000	0.639	0.000	0.000	0.000	0.056	0.111	0.000	0.000	0.111	0.000	0.056	0.083	0.000	0.194	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.639
0.722	0.250	0.056	0.000	0.000	0.000	0.611	0.111	0.000	0.000	0.056	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.167	0.000	0.056	0.000	0.000	0.250
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.083	0.000	0.111	0.000	0.000	0.111	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.056	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.639	0.000	0.056	0.639	0.000	0.000	0.000	0.000	0.167	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.083
0.111	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.111
0.722	0.000	0.000	0.750	0.083	0.639	0.000	0.750	0.000	0.000	0.194	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.111	0.000	0.000	0.056	0.806	0.000	0.000	0.000	0.000	0.000	0.083	0.000	0.000	0.111	0.000	0.000	0.000	0.000	1.000	0.278	0.000	0.000	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.111	0.000	0.194	0.083	0.111	0.778	0.000	0.056	0.000	0.611	0.000	0.000	0.556	0.000	0.000	0.000	0.056	0.000	0.000	0.806	0.000
0.000	0.000	0.000	0.083	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.056	0.000	0.000	0.000	0.000
0.000	0.000	0.000	0.056	0.639	0.361	0.194	0.000	0.000	0.139	0.000	0.000	0.750	0.000	0.000	0.000	0.083	0.056	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.000	0.833	0.750	0.000	0.000	0.000	0.000	0.111	0.083	0.000	0.000	0.000	0.000	0.000	0.000	0.639	0.000	0.000	0.000	0.000	0.056	0.000	0.000	0.000	0.000

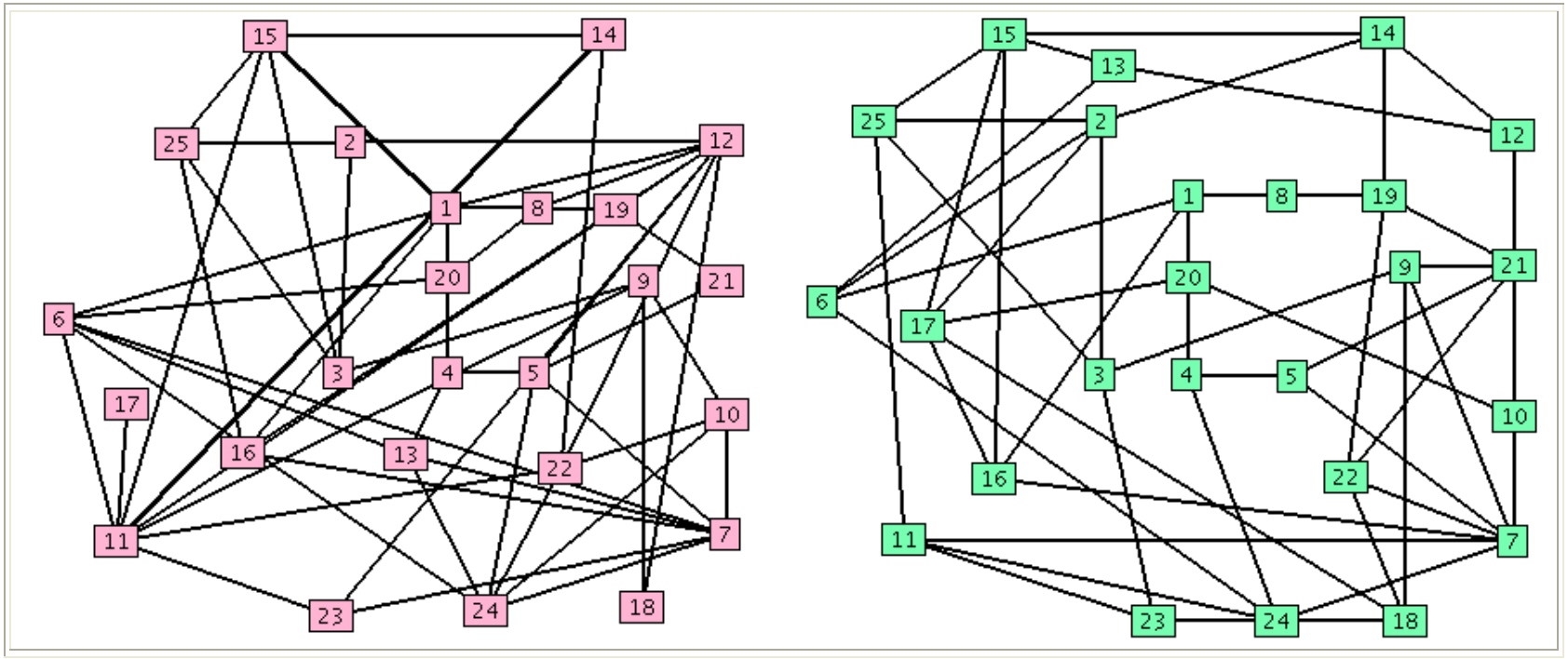


Figure 3.13: Graph from features matrix with 25 nodes from 5000 data points. Edges were included where at least 60% of sampled graphs had an edge.

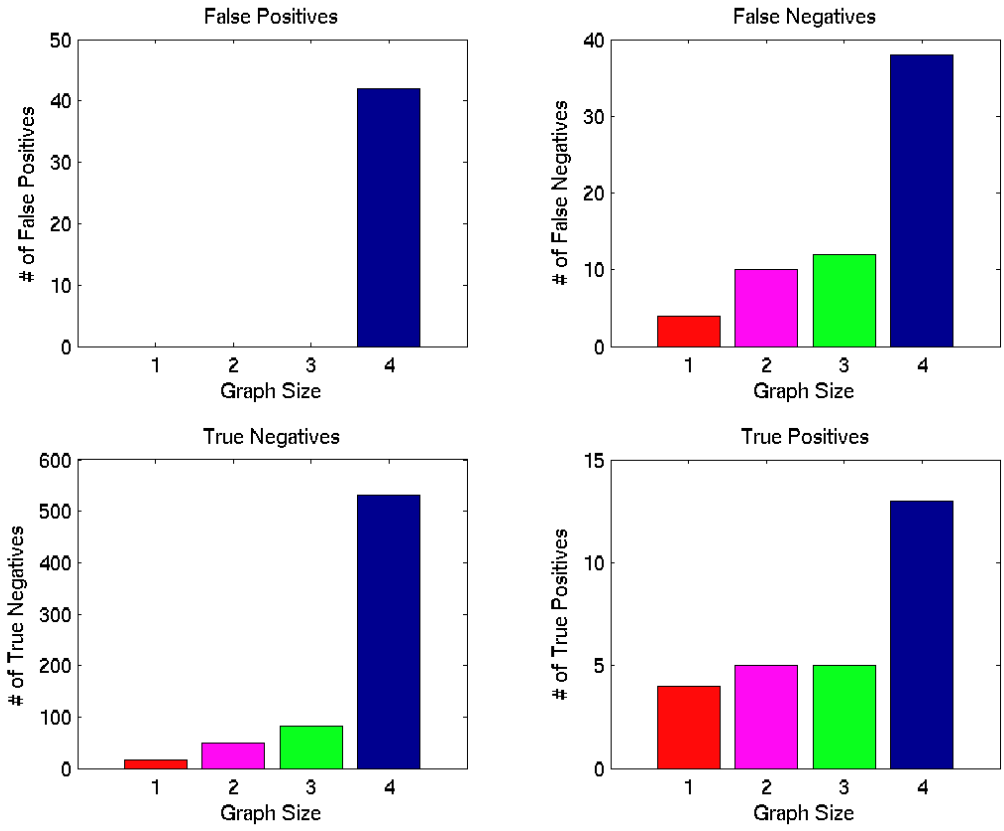


Figure 3.14: Scoring metrics of features graph with 5,000 data points. Blue line (Bar 4): 25 nodes, Green line (Bar 3): 10 nodes, Purple line (Bar 2): 8 nodes, Red line (Bar 1): 5 nodes.

V. Discussion

Two different simulations were completed in this chapter, which attempted to assess the accuracy of a sampling algorithm for learning the structure of a Bayesian Network. The first simulation tested the accuracy of learning the overall topological structure of a graph. Graphs of 5, 8, 10 and 25 nodes were tested, using an increasing amount of simulated data. One important problem in applying such methods to real biological data is that too much data is required to accurately learn the graph structure. This was tested by using graphs with known structures, and simulating data to represent them. Several different metrics were used to determine the accuracy of the learned graph versus the original graph. In the graph with 5 nodes, increasing the amount of data used to learn the graph seemed to have a positive effect on the learned graphs. The average number of true positives increased with the amount of data, although the number of true negatives did not seem to be affected. False positives actually increased slightly, while false negatives decreased, as was expected. Additionally, the number of empty rows present in the conditional probability tables decreased down to zero, proving that 5000 data points was a sufficient amount of data to learn the model. The same basic trends were seen in the graph with 8 nodes, with the exception that false positives decreased slightly, instead of increasing. Also there were still empty rows in the conditional probability tables, even with 5000 data points, which indicates that this is not enough data to learn the model. The graph with 10 nodes displayed the expected trends, where increasing the amount of data caused an increase in the number of true positives and true negatives, while decreasing the number of false negatives and false positives. The number of empty conditional probability tables decreased while the amount of data increased, although there was still an average of 3 empty rows, even with 5000 data points, which again indicates this is not a sufficient amount of data. When a graph with 25 nodes was tested, these same basic trends were observed, however there was a definite increase in the number of false positives and false negatives. Even with 5000 data points used, there was still an average of 27 false positives and 15 false negatives. This was a dramatic increase over these numbers in a graph of 10 nodes. Additionally, there was an average of more than 25 empty rows in the conditional probability tables with 5000 data points, again indicating a lack of sufficient data. With each of these four different sizes of graphs, the positive predictive value had an overall increase with the number of data points, while the false discovery rate had an overall decrease, although these were not perfect linear trends.

While it is important to consider the overall topological structure of a graph, as was done in the first simulation, it is also of interest to be able to identify nodes that are densely connected in the graph (i.e. have a high degree). This is particularly true in the application of such methods to the biological domain, as genes in a graph with a high degree would be of particular interest to test experimentally in knockout studies. Because of this, a second simulation was performed, which aimed to identify edges that are present in the majority of high-scoring graphs. This was tested in the same way as before, using four graphs of 5, 8, 10 and 25 nodes each, while supplying an increasing amount of data. Ten thousand graphs were sampled five times with each number of data points, and a different seed being used each time to initialize the random number

generator. Interestingly, it was found that increasing the amount of data beyond 300 points did not increase the highest scoring graph's score, nor did it increase the number of graphs that had a score within 3 of the highest scoring one. Since there were 10,000 graphs sampled in each iteration of the simulation, and 5 replicates, for every number of data points used for each size of graph, there were a total of 50,000 graphs sampled. Edges that were present in 60% of more of these 50,000 graphs were kept in a "features matrix", which is essentially an incidence matrix that contains percentages instead of a binary digit to indicate the presence or absence of an edge. From these matrices, graphs were built and visualized in ALIVE, to compare the original and learned graphs. While many edges were missing in the learned graphs, there were also many features that were correctly identified. No false positives were observed until the graph with 25 nodes was tested, in which there were more than 40 false positive edges present. The number of false negative edges in the learned graphs increased with node size, as did the number of true negative edges. The most important measure though, was the number of true positives. This had little increase between the graphs of 5, 8 and 10 nodes, although there was a dramatic increase in this number in the graph containing 25 nodes.

Overall, these results show that it might be possible to build Bayesian Networks from real biological data. While no model learned from the data could be considered to be exact, it could potentially identify the nodes in the graph with the highest degree (the hubs), which are of great interest in further biological dissection. It would be expected that many extraneous parts of the graph would be missed or incorrectly identified, however these are most likely of lesser biological importance anyways. The results above show though that using as few as 300 microarrays, a fairly accurate network could be learned from the data.

Works Cited

1. Radiation CtAHRFETLLOI. **Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2** 2005: [http://www.nap.edu/books/030909156X/html]
2. Rigaud O.: **The adaptive response to ionizing radiation: low dose effects unpredictable from high dose experiments.** *Hum Exp Toxicol* 1999, **18**(7):443-446.
3. Bonner WM.: **Low-dose radiation: thresholds, bystander effects, and adaptive responses.** *Proc Natl Acad Sci U S A* 2003, **100**(9):4973-4975.
4. Seong J KS, Pyo HR, Chung EJ, Suh CO.: **Low-dose radiation-induced adaptive response in the murine system.** *Ann N Y Acad Sci* 2002, **973**:255-257.
5. Sasaki MS EY, Tachibana A, Yamada T, Ishizaki K, Shimizu T, Nomura T.: **DNA damage response pathway in radioadaptive response.** *Mutat Res* 2002, **504**(1-2):101-118.
6. Takahashi A OK, Asakawa I, Kondo N, Nakagawa H, Yonezawa M, Tachibana A, Matsumoto H, Ohnishi T.: **Radiation response of apoptosis in C57BL/6N mouse spleen after whole-body irradiation.** *Int J Radiat Biol* 2001, **77**(9):939-945.
7. Lee SJ CS, Cho CK, Kim TH, Jeong KS, Yoo SY, Lee YS.: **Adaptive response is differently induced depending on the sensitivity to radiation-induced cell death in mouse epidermal cells.** *Cell Biol Toxicol* 2000, **16**(3):175-184.
8. Prasad KN CW, Hasse GM.: **Health risks of low dose ionizing radiation in humans: a review.** *Exp Biol Med* 2004, **229**(5):378-382.
9. Somers CM SR, Quinn JS, Boreham DR.: **Gamma radiation-induced heritable mutations at repetitive DNA loci in out-bred mice.** *Mutat Res* 2004, **568**(1):69-78.
10. Snyder AR MW: **Gene expression profiling after irradiation: clues to understanding acute and persistent responses?** *Cancer Metastasis Rev* 2004, **23**(2-4):259-268.
11. Yin E ND, Coleman MA, Peterson LE, Wyrobek AJ.: **Gene expression changes in mouse brain after exposure to low-dose ionizing radiation.** *Int J Radiat Biol* 2003, **79**(10):759-775.
12. Beck JA LS, Hafezparast M, Lennon-Pierce M, Eppig JT, Festing MF, Fisher EM.: **Genealogies of mouse inbred strains.** *Nat Genet* 2000, **24**(1):23-25.
13. Petkov PM DY, Cassell MA, Zhang W, Wagner G, Sargent EE, Asquith S, Crew V, Johnson KA, Robinson P, Scott VE, Wiles MV.: **An efficient SNP system for mouse genome scanning and elucidating strain relationships.** *Genome Res* 2004, **14**(9):1806-1811.
14. Yang YH DS, Luu P, Speed TP: **Normalization for cDNA Microarray Data**. In: *SPIE BiOS: 2001; San Jose, California*; 2001.
15. Yang YH DS, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation**. *Nucleic Acids Res* 2002, **30**(4):e15.
16. Kerr MK MM, Churchill GA: **Analysis of variance for gene expression microarray data**. *J Comput Biol* 2000, **7**(6):819-837.

17. Eckel JE GC, Therneau TM, Burgoon LD, Boverhof DR, Zacharewski TR: **Normalization of two-channel microarray experiments: a semiparametric approach**. *Bioinformatics* 2004.
18. Martin-Magniette ML AJ, Cabannes E, Daudin JJ.: **Evaluation of the gene-specific dye bias in cDNA microarray experiments.** *Bioinformatics* 2005, **21**(9):1995-2000.
19. Dobbin KK KE, Petersen DW, Simon RM.: **Characterizing dye bias in microarray experiments.** *Bioinformatics* 2005, **21**(10):2430-2437.
20. Smyth GK ST: **Normalization of cDNA microarray data**. *Methods* 2003, **31**(4):265-273.
21. Cohen RA. **An Introduction to PROC LOESS for Local Regression** 1999:
22. Dudoit S YY, Callow M, Speed T: **Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments**. In.; 2000: 38.
23. Fan J TP, Woude GV, Ren Y: **Normalization and analysis of cDNA microarrays using within-array replications applied to neuroblastoma cell response to a cytokine**. *Proc Natl Acad Sci U S A* 2004, **101**(5):1135-1140.
24. Yao B RS, Li Q, Ahmed S, Krauss R, Draghici S, Loeb JA: **Accuracy of cDNA microarray methods to detect small gene expression changes induced by neuregulin on breast epithelial cells**. *BMC Bioinformatics* 2004, **5**(1):99.
25. Oleksiak MF CG, Crawford DL: **Variation in gene expression within and among natural populations**. *Nat Genet* 2002, **32**(2):261-266.
26. Oleksiak MF RJ, Crawford DL: **Natural variation in cardiac metabolism and gene expression in Fundulus heteroclitus**. *Nat Genet* 2005, **37**(1):67-72.
27. Whitehead A CD: **Variation in tissue-specific gene expression among natural populations**. *Genome Biol* 2005, **6**(2):R13.
28. Rifkin SA KJ, White KP: **Evolution of gene expression in the Drosophila melanogaster subgroup**. *Nat Genet* 2003, **33**(2):138-144.
29. Jin W RR, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G: **The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster**. *Nat Genet* 2001, **29**(4):389-395.
30. Brem RB YG, Clinton R, Kruglyak L: **Genetic dissection of transcriptional regulation in budding yeast**. *Science* 2002, **296**(5568):752-755.
31. Pritchard CC HL, Delrow J, Nelson PS.: **Project normal: defining normal variance in mouse gene expression.** *Proc Natl Acad Sci U S A* 2001, **98**(23):13266-13271.
32. Turk R tHP, Sterrenburg E, de Menezes RX, de Meijer EJ, Boer JM, van Ommen GJ, den Dunnen JT.: **Gene expression variation between mouse inbred strains.** *BMC Genomics* 2004, **5**(1):57.
33. Cowles CR HJ, Altshuler D, Lander ES.: **Detection of regulatory variation in mouse genes.** *Nature Genetics* 2002, **32**(3):432-437.
34. Pavlidis P NW: **Analysis of strain and regional variation in gene expression in mouse brain.** *Genome Biol* 2001, **2**(10):RESEARCH0042.

35. Sandberg R YR, Pankratz DG, Carter TA, Del Rio JA, Wodicka L, Mayford M, Lockhart DJ, Barlow C: **Regional and strain-specific gene expression mapping in the adult mouse brain**. *Proc Natl Acad Sci U S A* 2000, **97**(20):11038-11043.
36. Cheung VG CL, Weber TM, Arcaro M, Jen KY, Morley M, Spielman RS: **Natural variation in human gene expression assessed in lymphoblastoid cells**. *Nat Genet* 2003, **33**(3):422-425.
37. Enard W KP, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, Doxiadis GM, Bontrop RE, Paabo S: **Intra- and interspecific variation in primate gene expression patterns**. *Science* 2002, **296**(5566):340-343.
38. Chesler EJ LL, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA, Threadgill DW, Manly KF, Williams RW.: **Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function.** *Nature Genetics* 2005, **37**(3):233-242.
39. Wade CM KEr, Kirby AW, Zody MC, Mullikin JC, Lander ES, Lindblad-Toh K, Daly MJ.: **The mosaic structure of variation in the laboratory mouse genome.** *Nature* 2002, **420**(6915):574-578.
40. Lee SG HJ, Kim YS: **A graph-theoretic modeling on GO space for biological interpretation of gene clusters.** *Bioinformatics* 2004, **20**(3):381-388.
41. Ashburner M BC, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G.: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nature Genetics* 2000, **25**(1):25-29.
42. Quackenbush J: **Computational analysis of microarray data.** *Nature Reviews Genetics* 2001, **2**(6):418-427.
43. Datta S DS: **Comparisons and validation of statistical clustering techniques for microarray gene expression data.** *Bioinformatics* 2003, **19**(4):459-466.
44. Tamayo P SD, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR.: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc Natl Acad Sci U S A* 1999, **96**(6):2907-2912.
45. Wang D RH, Musavi M, Domnisoru C.: **Double self-organizing maps to cluster gene expression data**. In: *European Symposium on Artificial Neural Networks: April 24-26 2002*; 2002: 45-50.
46. Ressom H WD, Natarajan P: **Clustering gene expression data using adaptive double self-organizing map.** *Physiol Genomics* 2003, **14**(1):35-46.
47. Xu Y OV, Xu D.: **Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees.** *Bioinformatics* 2002, **18**(4):536-545.
48. Tomida S HT, Honda H, Kobayashi T.: **Analysis of expression profile using fuzzy adaptive resonance theory.** *Bioinformatics* 2002, **18**(8):1073-1083.

49. Gasch AP EM: **Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering.** *Genome Biology* 2002, **3**(11):RESEARCH0059.
50. Shoemaker JS PI, Weir BS.: **Bayesian statistics in genetics: a guide for the uninitiated.** *Trends in Genetics* 1999, **15**(9):354-358.
51. Beaumont MA RB: **The Bayesian revolution in genetics.** *Nature Reviews Genetics* 2004, **5**(4):251-261.
52. Jordan MI.: **Graphical Models**. *Statistical Science* 2004, **19**:140-155.
53. Larget B SD: **Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees**. *Molecular Biology and Evolution* 1999, **16**(6):750-759.
54. Yang D ZS, Page GP, Brand JP, Edwards JW, Bartolucci AA, Allison DB.: **Applications of Bayesian statistical methods in microarray data analysis.** *American Journal of Pharmacogenomics* 2004, **4**(1):53-62.
55. Friedman N LM, Nachman I, Pe'er D.: **Using Bayesian networks to analyze expression data.** *Journal of Computational Biology* 2000, **7**(3-4):601-620.
56. Friedman N LM, Nachman I, Pe'er D. **Survey of Using Bayesian Networks to Analyze Expression Data** 2001: [http://math.berkeley.edu/~datta/]
57. Friedman N.: **Inferring cellular networks using probabilistic graphical models.** *Science* 2004, **303**(5659):799-805.
58. Kim SY IS, Miyano S.: **Inferring gene networks from time series microarray data using dynamic Bayesian networks.** *Briefings in Bioinformatics* 2003, **4**(3):228-235.
59. Jansen R YH, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M.: **A Bayesian networks approach for predicting protein-protein interactions from genomic data.** *Science* 2003, **302**(5644):449-453.
60. Nariai N KS, Imoto S, Miyano S.: **Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks.** In: *Pacific Symposium on Biocomputing: 2004*; 2004: 336-347.
61. Ghahramani Z.: **An Introduction to Hidden Markov Models and Bayesian Networks**. *International Journal of Pattern Recognition and Artificial Intelligence* 2001, **15**(1):9-42.
62. Woolf PJ PW, Daheron L, Daley GQ, Lauffenburger DA: **Bayesian analysis of signaling networks governing embryonic stem cell fate decisions.** *Bioinformatics* 2005, **21**(6):741-753.
63. Yu J SV, Wang P, Hartemink A, Jarvis E.: **Using Bayesian Network Inference Algorithms to Recover Molecular Genetic Regulatory Networks.** In: *International Conference on Systems Biology: December 2002*; 2002.
64. Ong IM GJ, Page D.: **Modelling regulatory pathways in E. coli from time series expression profiles.** *Bioinformatics* 2002, **18**(Supplement 1):S241-248.
65. Hartemink AJ GD, Jaakkola TS, Young RA.: **Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks.** *Pac Symp Biocomput* 2001:422-433.

66. Heckerman D GD, Chickering DM.: **Learning Bayesian Networks: The Combination of Knowledge and Statistical Data**. *Machine Learning* 1995, **20**(3):197-243.
67. Burnham KP AD: **Multimodel inference - understanding AIC and BIC in model selection**. *Sociological Methods & Research* 2004, **33**(2):261-304.
68. Koivisto M SK: **Exact Bayesian Structure Discovery in Bayesian Networks**. *Journal of Machine Learning Research* 2004, **5**:549-573.
69. Murphy KP: **The Bayes Net Toolbox for Matlab**. *Computing Science and Statistics* 2001.
70. Friedman N KD: **Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks.** *Machine Learning* 2003, **50**(1-2):95-125.
71. Yu J SV, Wang PP, Hartemink AJ, Jarvis ED.: **Advances to Bayesian network inference for generating causal networks from observational biological data.** *Bioinformatics* 2004, **20**(18):3594-3603.
72. Segal E FN, Koller D, Regev A.: **A module map showing conditional activity of expression modules in cancer.** *Nature Genetics* 2004, **36**(10):1090-1098.
73. Segal E WH, Koller D.: **Discovering molecular pathways from protein interaction and gene expression data.** *Bioinformatics* 2003, **19**(Suppl 1):i264-271.
74. Beal MJ FF, Ghahramani Z, Rangel C, Wild DL.: **A Bayesian approach to reconstructing genetic regulatory networks with hidden factors.** *Bioinformatics* 2004, **21**(3):349-356.
75. Segal E BY, Simon I, Friedman N, Koller D.: **From Promoter Sequence to Expression: A Probabilistic Framework**. *Proc 6th Inter Conf on Research in Computational Molecular Biology (RECOMB)* 2002.
76. Segal E SM, Regev A, Pe'er D, Botstein D, Koller D, Friedman N.: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.** *Nat Genet* 2003, **34**(2):166-176.
77. Sontag E KA, Kholodenko BN.: **Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data.** *Bioinformatics* 2004, **20**(12):1877-1886.
78. Yoo C TV, Cooper GF.: **Discovery of causal relationships in a gene-regulation pathway from a mixture of experimental and observational DNA microarray data.** *Pac Symp Biocomput* 2002:498-509.
79. Le PP BA, Ungar LH.: **Using prior knowledge to improve genetic network reconstruction from microarray data**. *In Silico Biology* 2004, **4**(1):335-353.
80. van Dijk S vdGL, Thierens D.: **A skeleton-based approach to learning Bayesian networks from data.** *Lecture Notes in Computer Science* 2003, **2838**.
81. Ott S HA, Kim SY, Miyano S: **Superiority of network motifs over optimal networks and an application to the revelation of gene network evolution**. *Bioinformatics* 2005, **21**(2):227-238.
82. Pe'er D RA, Elidan G, Friedman N.: **Inferring subnetworks from perturbed expression profiles.** *Bioinformatics* 2001, **17**(Supplement 1):S215-224.

83. Wessels LF vSE, Reinders MJ: **A comparison of genetic network models**. *Pac Symp Biocomput* 2001:508-519.
84. Rinchik G.: **DOE LD-RD Grant Application**. In.; 2001.
85. Wolfinger RD GG, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS: **Assessing gene significance from cDNA microarray expression data via mixed models**. *J Comput Biol* 2001, **8**(6):625-637.
86. Thygesen HH ZA: **Comparing transformation methods for DNA microarray data.** *BMC Bioinformatics* 2004, **5**(1):77.
87. Cui X CG: **Statistical tests for differential expression in cDNA microarray experiments**. *Genome Biol* 2003, **4**(4):210.
88. Pollard KS vdLM. **Cluster Analysis of Genomic Data with Applications in R** 2005: [http://www.bepress.com/ucbbiostat/paper167]
89. Audit B OC: **From genes to genomes: universal scale-invariant properties of microbial chromosome organisation.** *Journal of Molecular Biology* 2003, **332**(3):617-633.
90. Bisognin A BS, Danieli GA.: **Detection of chromosomal regions showing differential gene expression in human skeletal muscle and in alveolar rhabdomyosarcoma.** *BMC Bioinformatics* 2004, **5**(1):68.
91. Caron H vSB, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA, Heisterkamp S, van Kampen A, Versteeg R.: **The human transcriptome map: clustering of highly expressed genes in chromosomal domains.** *Science* 2001, **291**(5507):1289-1292.
92. Kendal WS.: **A scale invariant clustering of genes on human chromosome 7.** *BMC Evol Biol* 2004, **4**(1):3.
93. Kirov SA PX, Baker E, Schmoyer D, Zhang B, Snoddy J: **GeneKeyDB: a lightweight, gene-centric, relational database to support data mining environments.** *BMC Bioinformatics* 2005, **6**(1):72.
94. Zhang B SD, Kirov S, Snoddy J: **GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies.** *BMC Bioinformatics* 2004, **5**(1):16.
95. Pavlidis P LQ, Noble WS: **The effect of replication on gene expression microarray experiments.** *Bioinformatics* 2003, **19**(13):1620-1627.
96. Hartemink AJ GD, Jaakkola TS, Young RA.: **Combining location and expression data for principled discovery of genetic regulatory network models.** *Pac Symp Biocomput* 2002:437-449.
97. Giudici P CR: **Improving Markov Chain Monte Carlo model search for data mining**. *Machine Learning* 2003, **50**(1-2):127-158.
98. Aiello W CF, Lu L.: **A random graph model for massive graphs**. In: *Proceedings of the thirty-second annual ACM symposium on Theory of computing: 1999; Portland, Oregon, United States*: ACM; 1999: 171 - 180.
99. Palmer CR SJ: **Generating Network Topologies That Obey Power Laws**. In: *Proceedings of the Global Internet Symposium, Globecom2000: 2000*; 2000.
100. Barabasi AL OZ: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 2004, **5**(2):101-113.

101. Basso K MA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: **Reverse engineering of regulatory networks in human B cells**. *Nat Genet* 2005, **37**(4):382-390.
102. Song C HS, Makse HA: **Self-similarity of complex networks**. *Nature* 2005, **433**(7024):392-395.
103. Schafer J SK: **An empirical Bayes approach to inferring large-scale gene association networks.** *Bioinformatics* 2005, **21**(6):754-764.

Vita

I graduated from Saint Francis University in Loretto, Pennsylvania in May of 2002, with a bachelor's of science degree. In pursuing this degree, I double majored in computer science and biology, finishing with a 3.595 grade point average, on a scale of 4.0. While at Saint Francis, I participated in an undergraduate research project for a year, which studied the effects of sir3 as a potential silencer and enhancer in yeast. I also interned for a summer at Purdue Pharma, L.P., where I worked in the pharmacokinetics and drug metabolism division. In this position, I assisted a primary scientist with drug metabolism, inhibition and induction research, using human hepatocytes as a culture medium. Effects of the drugs tested were analyzed using high performance liquid chromatography. In this position, I also gained experience with hepatocyte and microsomal incubations and isolation.

Immediately following my graduation from Saint Francis, I entered graduate school at the University of Tennessee/Oak Ridge National Laboratory's Graduate School of Genome Science and Technology. In my first year of this program I participated in four research rotations, within the ORNL genome analysis & systems modeling group, ORNL mammalian genetics & genomics group, and the UT department of nutrition. In these rotations, I worked with RNA isolation, microarray hybridizations, data analysis and enzymatic assays. I also designed web interfaces, extended development of open-source projects and worked on computational grid development. Following my research rotations, I spent two years of research on my master's thesis. During this time, I analyzed microarray data, which included normalization, differential expression, analysis of variance, clustering, cluster analysis and network construction. I also applied methods for learning Bayesian Networks from high throughput experimental data. In addition, I created a Java applet for the visualization of network topology and graph characteristics, as well as developed databases and dynamic web interfaces for the sharing of data and results for collaborative scientific research projects. While at UT, I also worked for the student newspaper, the Daily Beacon, by maintaining, managing and developing the paper's website. I also created a student events calendar, custom web interface and content management system for use by the campus and local community.