CHALMERS

Comparing Inferred Preferential Gene-Tissue Relationships in Human and Other Species

Master of Science in Bioinformatics and Systems Biology

Padmanabhuni Shanmukha Sampath

Supervisors: Daniel Dalevi
Marcus Bjäreland
R&D Information
AstraZeneca
SE-431 83 Mölndal
Sweden

Examiner: Olle Nerman
Department of Mathematical Science
Chalmers University of Technology
SE-412 96 Göteborg
Sweden

Department of Mathematical Science
Chalmers University of Technology
Göteborg, Sweden, SE-412 96
February 2012

Comparing Inferred Preferential Gene-Tissue Relationships in Human and Other Species

PADMANABHUNI SHANMUKHA SAMPATH

Department of Mathematical Science
Chalmers University of Technology
Göteborg, Sweden, SE-412 96
February 2012

Abstract

In drug discovery a gene expressed only in one or a few tissues is of more interest than a gene ubiquitously expressed in many tissues. As testing of drugs normally is initiated in animal models before humans, a computational approach to identify tissue-specific genes across several species is highly relevant. In this thesis we aim to identify tissue-specific (special case of tissue-selective, selectively expressed in one tissue), 2-selective (tissue-selective, preferentially expressed in two tissues), 3-selective (tissue-selective, preferentially expressed in three tissues) and 4-selective (tissue-selective, preferentially expressed in four tissues) genes across Human, Mouse and Rat datasets.

We have collected five diverse datasets (three Human, one Mouse and one Rat) to identify tissue-selective genes. All of these datasets have replicates which enabled the use of a Bayesian factor method. We have also collected training data for all three species. The Bayesian factor method was extended, parameters were optimized and it was applied on all datasets.

The results were satisfactory except for one noisy Human dataset. The overlap in the identified tissue-specific genes between the Human and Mouse species is low, 10%, and between the Human and Rat species is very low, 4%. The overlap in the identified tissue-specific genes between the Mouse and Rat species is 3.1%. The low amount of overlap in tissue-specific genes across species is due to insufficient amount of training genes.

Acknowledgement

This project was carried out within Research and Development information at AstraZeneca Mölndal. It was my pleasure working in this project which gave me a lot of good experiences and enriched my knowledge in gene expression, next generation sequencing and statistical methods and tools. I have also gained experience working in pharmaceutical industrial environment.

I would like to express my deepest gratitude to my supervisor Daniel Dalevi for his wonderful guidance throughout the project and for all the wonderful discussions we had. I would like to thank Daniel Dalevi for his ideas and encouragement he has given me during the project. I would like to express my gratitude to Marcus Bjäreland for his guidance in writing thesis report and for nice discussions we had. I have no words to express my gratitude to both of them. I would like thank Olle Nerman from Chalmers for all the wonderful discussions we had regarding the statistical methods and for his ideas which gave us confidence in doing the project.

I would like to thank Saber Ahmad Akhondi and Jing Guo for their help and support during the project. I would like to thank Selam Tesfaye for being a wonderful colleague. I would like to thank Lokeshwaran Manoharan for helping me with structure of the thesis report. I would like to thank Liu Peidi for letting me know about this project. I would like to thank all my classmates and friends for the support. I would like to thank Mariana Pereira for reviewing and commenting my thesis report.

Finally, I would like to thank my family, especially my father Sampath Kumar Papdmanbhuni for his teachings using real world analogies.

Table of Contents

1 Introduction

The thesis is done in AstraZeneca at Mölndal.

1.1 Background

The identification of preferential Gene-Tissue relationships is an important research topic to explore in:

1. Disease prognosis.
2. Validation of tissue-specific targets.
3. Evolutionary processes.

Biomarkers are used in disease prognosis to monitor the disease. Biomarkers are measured in blood plasma or serum and the precise source is an important factor for the measurement. Information regarding Gene-Tissue relationships can help in selection of suitable biomarkers to ensure that analytes are from the right source tissue. The probability of a tissue-specific gene to become a drug target is high compared to genes that are ubiquitously expressed in many tissues (Dezso, Nikolsky et al. 2008). Selection of potential drug targets is an important task in drug discovery. Transcriptome analysis is performed on whole genomes to measure gene expression across multiple tissues to predict tissue-specific genes. Transciptome analysis is also helpful in finding relationships between a protein's expression pattern and its evolutionary origin or functional category (Freilich, Massingham et al. 2005).

Gene-Tissue relationships can be estimated from quantitative data measuring gene (or protein) expression, such as Microarrays, Expressed Sequence Tags (EST), Massively Parallel Signature Sequencing (MPSS) and RNA-Seq. There are several methods to predict tissue-specific genes, some are: ROKU (Kadota, Ye et al. 2006), SPM (Xiao, Zhang et al. 2010), Bayesian Factor (Van Deun, Hoijtink et al. 2009) and Tukey-Kramer test (Liang, Li et al. 2006). A computational approach was developed to identify preferential Gene-Tissue relationships by combining results from several methods/datasets and was applied only on Human data. The fact that testing of drugs is done on animals first before humans gives an important task to estimate if a gene is tissue-specific across a set of species used in pharmaceutical trials like Mouse (GEO Accession: GSE9954), Rat (GEO Accession: GSE952).

Different types of Gene-Tissue relationships are (Van Deun, Hoijtink et al. 2009):

1. Tissue-specificity (gene expressed only in one tissue),
2. Categorical tissue-specificity (gene differently expressed in one tissue compared to the rest),
3. Tissue-selectivity (gene differently expressed in only one or a few tissues) and
4. Ubiquitously expressed (gene expressed in all or many tissues).

The first and the second types of Gene-Tissue relationships define tissue-specific genes. Tissue-specificity is a special case of the categorical tissue-specificity. The categorical tissue-specificity is a special case of tissue-selectivity. The third type of Gene-Tissue relationships defines tissue-selective genes and the fourth type defines ubiquitously

expressed genes. Less number of tissue-specificity (gene expressed only in one tissue) genes make most of the research to be carried out on tissue-selectivity and categorical tissue-specificity genes.

In this project we looked for tissue-specific genes and tissue-selective genes. Tissue-specificity is a special case of tissue-selectivity where a gene is selectively expressed in only one tissue. The definition of tissue-selectivity was vague as it does not explain how many tissues are few tissues. After looking at training data (Section 2.2), collected for all species, we defined few tissues to be less than or equal to four. Tissue-selectivity hence means that a gene is expressed in one, two, three or four tissues. Therefore the different types of genes we explore in this project are (Figure 1):

1. Tissue-specific genes e.g. CASP14 (selectively expressed in one tissue only),
2. 2-selective genes e.g. INPP5F (selectively expressed in two tissues only),
3. 3-selective genes e.g. AMY2B (selectively expressed in three tissues only) and
4. 4-selective genes e.g. PIGR (selectively expressed in four tissues only).

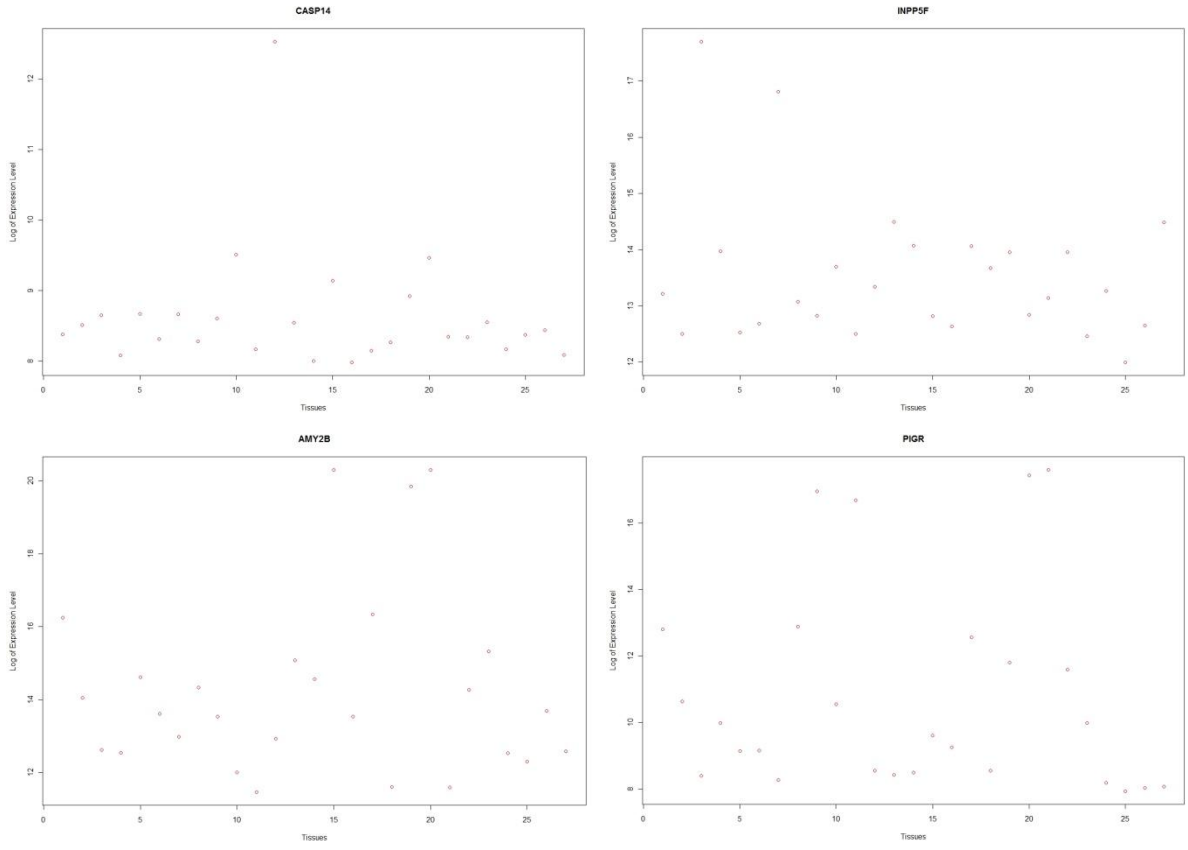


Figure 1 The expression pattern of a tissue-specific gene (upper left), a 2-selective gene (upper right), a 3-selective gene (lower left) and a 4-selective gene (lower right).

If a gene does not belong to any of the above mentioned types, it will be assigned as ubiquitously expressed. Selective gene expression can occur in two directions: up regulated (gene is over-expressed in one or a few tissues compared to other) or down regulated (gene is under-expressed in one or a few tissues compared to the others). In this project we only study up regulated genes because they are more relevant in pharmaceutical industry (Dezso, Nikolsky et al. 2008).

We focused on the Bayesian factor method (Van Deun, Hoijtink et al. 2009) and the marginal likelihood estimators (Chib 1995) in order to predict tissue-specific genes (Section 3.1.1). We have extended (Section 3.1.2 - 3.1.4) and modified (Section 3.2) the Bayesian factor method to identify tissue-selective genes. The Bayesian factor approach requires datasets with replications to be able to detect either tissue-specific or tissue-selective genes. Before we could apply the method we worked on preparing the datasets (Section 2.1) since they differed in terms of tissues, replicates, species and strains. We have done vocabulary mapping (Section 3.4.1) and grouping (Section 3.4.2) on the datasets to make results from different datasets comparable. For the Rat (GSE952) dataset we have calculated the ratio between variance between strains and variance within strains (Section 3.5) to see whether we can combine the data from different strains or not. We ran the Bayesian factor method, after optimizing the parameters, on the datasets and the results are discussed (Section 4).

1.2 Organization

The aim of this project is to predict tissue-specific, 2-selective, 3-selective and 4-selective genes in the Human, Mouse and Rat species. The organization of this project is as follows:

1. The datasets with replicates are selected and fixed.
2. The Bayesian factor method is modified and extended.
3. The parameters for the Bayesian factor method are optimized in each dataset using training genes.
4. The Bayesian factor method is applied to all datasets with optimized parameters to estimate specific and selective genes.
5. Overlap of the resulting tissue-specific genes among the datasets is calculated.

1.3 Outline

The Data section (Section 2) provides information on the datasets, training genes and test genes used in the project. The Method section (Section 3) explains the Bayesian factor, extension and modification of the Bayesian factor and optimizing parameters. The Result section (Section 4) explains results from the Bayesian factor method on each dataset. The Discussion and Conclusion section (Section 5) explains the future work of the project.

2 Data

The Bayesian factor method works only for data with replicates. We collected five datasets which are available in GEO (Gene Expression Omnibus). Training and testing genes were collected for all species.

2.1 Datasets

All five datasets have replicates as the Bayesian factor method works only for data with replicate samples. All five datasets contain expression level of various tissues in a gene. The number of tissues covered and number of replicates per tissue varies among the datasets. The datasets are:

1. GDS3113: The dataset is on expression of various normal tissues in humans (Dezso, Nikolsky et al. 2008). There are 32878 probesets out of which 18002 probesets are mapped to genes and 14876 unmapped probesets. Before vocabulary mapping (Section 3.4.1) the dataset covers 32 tissues which are grouped to 27 tissues. Each tissue has 3 replicates.

2. GDS596: The dataset is on expression of normal Human tissues (Su, Wiltshire et al. 2004). There are 22283 probesets in the dataset out of which 21169 probesets are mapped to genes. The dataset covers 79 tissues before vocabulary mapping (Section 3.4.1) and 30 tissues after vocabulary mapping. Each tissue has 2 replicates. The dataset has high noise level in the data.

3. GSE7307: The dataset is on expression of normal Human tissues covering 105 tissues. After vocabulary mapping (Section 3.4.1) the dataset covers 42 tissues each having different replicates. There are 54675 probesets with 42369 mapped probesets and 12306 unmapped probesets.

4. GSE9954: The dataset is on expression of various Mouse tissues (Lieven Thorrez 2008). The dataset cover 22 tissues before vocabulary mapping (Section 3.4.1) and 19 tissues after vocabulary mapping. Out of 45101 probesets in the dataset 39331 probesets are mapped to genes and 5770 probesets are not mapped. Most of the tissues have 3 replicates except for a few which have 4 or 5 replicates.

5. GSE952: The dataset is on transcriptome analysis in the Rat (Walker, Su et al. 2004). This is the smallest dataset out of all the five dataset with 8799 probesets. Out of 8799 probesets 7722 probesets are mapped to genes. Three albino strains of Rat are used in the experiment namely Wistar, Wistar Kyoto and Sprague Dawley. The dataset covers 26 tissues grouped to 12 using vocabulary mapping (Section 3.4.1). Most of the tissues have 2 replicates except for one tissue which has 6 replicates.

2.2 Training and Test Data

Training data was not needed for the original implementation, i.e. specificity, as the threshold could be chosen from the parameters of the dataset. After our modifications

training data is needed to optimize the parameters as we have introduced a new parameter for the variability between tissues. Training data is collected for all datasets from papers (Hu, Chen et al. 2001; Knoll, Pietrusz et al. 2005; Kouadjo, Nishida et al. 2007; Stansberg, Vik-Mo et al. 2007; Silver, Cotroneo et al. 2008). The genes covered in training data were mostly tissue-specific genes and some 2-selective genes for all species. The quantities of 3-selective and 4-selective genes covered are very less in training data for all species. Test data is collected from the TISGED (Xiao, Zhang et al. 2010) database for all species.

3 Method

3.1 Bayesian Factor and Extended Bayesian Factor

The method explained is in agreement with (Van Deun, Hoijtink et al. 2009) and defined for one gene. In their (Van Deun, Hoijtink et al. 2009) notation and formulation, the data consist of $i = 1, \dots, N$ expression levels y_i for $j = 1, \dots, J$. The total sample size $N=\sum_j N_j$ with N_j denoting the number of replications for tissue j which means that i is nested in j. So, for example, for $j = 1$, $i = 1,\dots,N_1$, for $j = 2$, $i = N_1+1, \dots, N_1+N_2$. The model for the gene expression level is

$$y_i = \mu_1 d_{1i} + \mu_2 d_{2i} + \cdots + \mu_J d_{Ji} + \varepsilon_i,$$

where μ_j denotes the population mean of tissue j, d_{ji} is 1 if the expression was obtained for tissue j else it is 0 and $\varepsilon_i \sim N(0,\sigma^2)$. The density of the data for this ANOVA model is

$$f\left(y|d_1, \dots, d_J, \mu_1, \dots, \mu_J, \sigma^2\right) = \prod_j \prod_{i \in j} \frac{1}{\sqrt{2\pi\sigma^2}} exp - \frac{1}{2} \frac{(y_i - \mu_1 d_{1i} - \cdots - \mu_J d_{Ji})^2}{\sigma^2},$$

where $y = [y_1, \dots, y_N]$ and $d_j = [d_{j1}, \dots, d_{jN}]$.

The Bayesian factor method in (Section 3.1.1) is defined only for the tissue-specific case (Section 1.1). From (Section 3.1.2 - 3.1.4) extended method which covers the other gene expression patterns but the algorithm is the same as explained in tissue-specific case (Section 3.1.1).

3.1.1 Tissue-Specific Case

In connection to model comparisons the prior probabilities π_1 and π_2 of two models M_1 and M_2, both models are Bayesian, for the parameters in these distributions are given conditional on which of the two models that are true. This give rise to a complete model which can be seen as an overall Bayesian model M_u with a prior distribution $h(\dots)$. The unconstrained hypothesis

$$H_u : \mu_1, \dots, \mu_J,$$

which is the nested model of constrained hypotheses H_m where $m=1,\dots,T$ with different number and type of constraints. As a result of the nesting only prior distribution of unconstrained hypothesis has to be specified which is denoted by $h(\theta^u, \sigma^2|H_u)$ where $\theta^u = [\mu_1, \dots, \mu_J]$. The prior distribution of unconstrained hypothesis H_u (Irene Klugkist 2007) is

$$h(\theta^u, \sigma^2|H_u) = \prod_j \mathrm{N}(\mu_j|\mu_0, \tau_0)\mathrm{Inv}\text{-}\chi^2(\sigma^2|\nu, \sigma_0^2),$$

where choice of μ_0, τ_0, σ_0^2 and ν is independent indicating that prior distribution of each μ_j is same and the notation $\mathrm{N}(\dots)$ and $\mathrm{Inv}\text{-}\chi^2(\dots)$ stands for densities. The aim is to look for tissue-specific genes in this case so two hypotheses were considered:

$$H_1 : \mu_1 > \{\mu_2, \ldots, \mu_J\},$$

which states that μ_1 is larger than each of the mean in the set $\{\mu_2, \ldots, \mu_J\}$, and

$$H_2 : \text{not } H_1.$$

μ_1 is the target tissue mean. The parameter space of hypothesis H_1 is Θ_1 and for hypothesis H_2 is Θ_2. In choosing between the two hypotheses H_1:$\theta \in \Theta_1$ and its complement in a classical Bayesian model the choice may be interpreted as choosing between two submodels M_1, which is identified with the conditional model got from constraints on $\theta \in \Theta_1$, and M_2 the conditional model got from conditioning on $\theta \in \Theta_2$ which is compliment of Θ_1. The prior distribution of a constrained hypothesis H_m with parameter $\theta^m \in \Theta_m$ follows directly from the prior distribution of unconstrained hypothesis H_u, using

$$\begin{aligned} h(\theta^m, \sigma^2|H_{\mathrm{m}}) &= \frac{h(\theta^m, \sigma^2|H_u) I_{H_{\mathrm{m}}}}{\int_{\theta^u, \sigma^2} h(\theta^m, \sigma^2|H_u) I_{H_{\mathrm{m}}} d\theta^m d\sigma^2} \\ &= c_m h(\theta^m, \sigma^2|H_u), \end{aligned}$$

where I_{H_m} is 1 if parameter values are in agreement with the constraints of the hypothesis H_m and 0 otherwise, and c_m is proportion of H_u in agreement with H_m. Similarly the posterior distribution of hypothesis H_m with parameter $\theta^m \in \Theta_m$ can be written in terms of posterior distribution of H_u

$$\begin{aligned} g(\theta^m, \sigma^2|y, H_{\mathrm{m}}) &= \frac{g(\theta^m, \sigma^2|y, H_u) I_{H_{\mathrm{m}}}}{\int_{\theta^u} g(\theta^m, \sigma^2|y, H_u) I_{H_{\mathrm{m}}} d\theta^m} \\ &= f_m g(\theta^m, \sigma^2|y, H_u), \end{aligned}$$

where f_m is proportion H_u in agreement with H_m. The Bayes factor for choosing between the two conditional models M_1 and M_2 with hypotheses H_1 and H_2 in this parameter setting becomes the odds ratio:

$$BF_{12} = \frac{P(y|H_1)}{P(y|H_2)} = \frac{P_g(\theta \in \Theta_1, \sigma^2 \,|\mathrm{y}, H_1)/P_h(\theta \in \Theta_1, \sigma^2|H_1)}{(1 - P(\theta \in \Theta_1, \sigma^2 \,|\mathrm{y}, H_1))\,/(1 - P_h(\theta \in \Theta_1, \sigma^2|H_1))} = \frac{c_2 f_1}{c_1 f_2},$$

where P_g is probability of the posterior distribution and P_h is probability of prior distribution. BF_{12} = 12 indicates that the support in data for H_m is 12 times as large as the support for hypothesis H_k. Under the specifications for unconstrained model prior distribution, $c_1 = 1/J$ and c_2=1-c_1 = (J-1)/J, as there are J equivalent models in which one of the means if larger than other means. Substituting the prior and posterior distributions of the hypothesis H_1 and H_2 the Bayesian factor BF_{12} is

$$BF_{12} = \frac{BF_{1\mathrm{u}}}{BF_{2u}} = \frac{c_2 f_1}{c_1 f_2} = \frac{J f_1}{\frac{J}{(J-1)} f_2} = (J-1)\frac{f_1}{f_2}.$$

Given c_1 and c_2 all that is left is to calculate f_1 as $f_2 = 1 - f_1$. Any number for μ_0, degree of freedom $\nu = 2$, very large τ_0 and $\sigma_0^2 = 0$ meaning vague priors are used to estimate f_1.

Estimation of f_1 from the unconstrained posterior distribution is done by following steps of the algorithm (Gibbs Sampling):

1. Assign initial values: $\mu_j = \bar{y}_j$ for $j = 1, \ldots, J$ and
$$\sigma^2 = \frac{1}{N}\sum_i (y_i - \bar{y}_1 d_{1i} - \cdots - \bar{y}_J d_{Ji})^2,$$
where $\bar{y}_j$ denotes the sample average for tissue j.

2. For $j = 1, \ldots, J$ sample μ_j from $g(\mu_j|\sigma^2,\text{y})$ which is a normal distribution with mean $\bar{y}_j$ and variance σ^2/N_j.
3. Check whether the tissue means $\mu_1,\ldots,\mu_J$ are in agreement with constraints of hypothesis H_1.

4. Sample the variance σ^2 from $\text{g}(\sigma^2|\mu_1,\ldots,\mu_J, \text{y})$ according to the inverse χ^2 distribution with N-2 degrees of freedom and scale parameter is
$$\frac{1}{(N-2)}\sum_i \left(y_i - \mu_i d_{1i} - \ldots - \mu_J d_{Ji}\right).$$

5. Repeat steps 2 to 4 to count number of times the target tissue is greater than the rest of the tissues in a gene and divide it by number of times the variance is sampled which will give an estimate of f_1.

As the prior distributions are uninformative, posterior distribution is proportional to likelihood of data meaning that estimation of f_1 depends of the data. Considering priors are same for all tissues indicates that c_1 and c_2 does not depend on prior distribution. In (Irene Klugkist 2007) it is argued that often improper uninformative priors can be used in a natural way in these kinds of situations when the hypothesis concern symmetrical inference situations with inequalities simply by using the natural interpretation of the model on the joint model level. Increase in the number of tissues smaller than the target tissue indicates that the Bayesian factor is in support of hypothesis H_1 and decrease in number of tissues less than target tissue indicates that the Bayesian factor is in support of hypothesis H_2. If all tissue means are equal indicates that the Bayesian factor is neutral with respect to both hypothesis which means $BF_{12} = 1$. The whole process is only done to the tissue with largest mean in a gene in order to avoid unnecessary computations.

3.1.2 2-Selective Case

The hypotheses to be tested for the 2-selective case are

$$H_1 : \min(\mu_1, \mu_2) > \{\mu_3, \ldots, \mu_J\},$$
$$H_2 : \text{not } H_1,$$

where μ_1 and μ_2 are the target tissue means. If the minimum of μ_1 and μ_2 is greater than the rest of the tissue means then the maximum of μ_1 and μ_2 is also greater than the rest of the tissue means in a gene. The calculation of Bayesian factor for H_1 against H_2 is same as explained in the (Section 3.1.1) with some modifications in the calculation of c_1 and c_2. There are two target tissues larger than rest of tissues which makes $c_1 = \frac{2}{J(J-1)}$

and $c_2 = \frac{(J-2)(J+1)}{J(J-1)}$ as $c_2 = 1 - c_1$. The algorithm to estimate f_1 is same as mentioned in (Section 3.1.1). The Bayesian factor BF_{12} is

$$BF_{12} = \frac{(J-2)(J+1)}{2}\frac{f_1}{f_2}.$$

3.1.3 3-Selective Case

The hypotheses to be tested for 3-selective case are

$$H_1 : \min(\mu_1, \mu_2, \mu_3) > \{\mu_4, \dots, \mu_J\},$$
$$H_2 : \text{not } H_1,$$

where μ_1, μ_2 and μ_3 are the target tissue means. The calculation of Bayesian factor for H_1 against H_2 is similar to the way it is calculated in the (Section 3.1.1) except for the calculation of c_1 and c_2. In this case there are three target tissues greater than rest of tissues which gives $c_1 = \frac{6}{J(J-1)(J-2)}$ and $c_2 = \frac{(J-3)(J^2+2)}{J(J-1)(J-2)}$ as $c_2 = 1 - c_1$. The Bayesian factor BF_{12} is

$$BF_{12} = \frac{(J-3)(J^2+2)}{6}\frac{f_1}{f_2}.$$

3.1.4 4-Selective Case

The hypotheses to be tested for 4-selective case are

$$H_1 : \min(\mu_1, \mu_2, \mu_3, \mu_4) > \{\mu_5, \dots, \mu_J\},$$
$$H_2 : \text{not } H_1,$$

where μ_1, μ_2, μ_3 and μ_4 are the target tissue means. The calculation of Bayesian factor for H_1 against H_2 is similar to the way it is calculated in the (Section 3.1.1). c_1 and c_2 change as there are four target tissues greater than rest of tissues. So $c_1 = \frac{24}{J(J-1)(J-2)(J-3)}$ and $c_2 = \frac{(J-4)(J^3-2J^2+3J+6)}{J(J-1)(J-2)(J-3)}$ as $c_2 = 1 - c_1$. The Bayesian factor BF_{12} is

$$BF_{12} = \frac{(J-4)(J^3-2J^2+3J+6)}{24}\frac{f_1}{f_2}.$$

3.2 Modified Bayesian Factor

The Bayesian factor and its extension explained in (Section 3.1) resulted in support for hypothesis H_1 for most of the genes because the standard deviation of each tissue is very small compared to the standard deviation between tissues in a gene. Most of the genes were shown as tissue-specific genes. The hypotheses to be tested on the data are modified in order to include the variability between tissues in a gene. The formulation of the hypotheses is modified by addition of extra parameters. The modified hypotheses would also take variability between tissues into account in the Bayesian factor calculation. The estimation of f_m for a hypothesis H_m is still same as explained in (Section 3.1.1).

3.2.1 Tissue-Specific Case

The modified hypotheses to be tested on the data are

$$H_1 : \mu_1 > \mu_{2\ldots J}^{max} + CS_{2\ldots J} ,$$
$$H_2 : \text{not } H_1 ,$$

where C is a constant, $S_{2\ldots J}$ is the sample standard deviation of tissues 2 to J, μ_1 is the maximum mean over all tissues in a gene and $\mu_{2\ldots J}^{max}$ is the maximum mean over 2 to J. The constant C varies for different datasets, types of tissue selectivity and for training data. The constant C is chosen to fit training data for respective datasets. The constant C is the parameter that includes the variability between the tissues in the hypothesis in order to avoid false positive cases in the data. In this situation the prior probabilities are not as direct in the (Section 3.1.1) as there is a hypothesis scale and translational constraint. The prior probability of H_1 turns out to be independent of the parameters in the prior distribution and one can simply calculate prior probabilities for a finite choice and use that also for the improper priors. It can be observed that the priors are still disjoint but they do not add to one anymore.

3.2.2 2-Selective Case

The modified hypotheses to be tested for 2-selective case are

$$H_1 : \min(\mu_1, \mu_2) > \mu_{3\ldots J}^{max} + CS_{3\ldots J} ,$$
$$H_2 : \text{not } H_1 ,$$

where μ_1 and μ_2 are the largest means over all tissues, $S_{3\ldots J}$ is the sample standard deviation of tissues 3 to J and $\mu_{3\ldots J}^{max}$ is the maximum mean over tissue 3 to J.

3.2.3 3-Selective Case

The modified hypotheses to be tested for the 3-selective case are

$$H_1 : \min(\mu_1, \mu_2, \mu_3) > \mu_{4\ldots J}^{max} + CS_{4\ldots J} ,$$
$$H_2 : \text{not } H_1 ,$$

where μ_1, μ_2 and μ_3 are the largest means over all tissues, $S_{4\ldots J}$ is the sample standard deviation of tissues 4 to J and $\mu_{4\ldots J}^{max}$ is the maximum mean over tissue 4 to J.

3.2.4 4-Selective Case

The modified hypotheses to be tested for the 4-selective case are

$$H_1 : \min(\mu_1, \mu_2, \mu_3, \mu_4) > \mu_{5\ldots J}^{max} + CS_{5\ldots J} ,$$
$$H_2 : \text{not } H_1 ,$$

where μ_1, μ_2, μ_3 and μ_4 are the largest means over all tissues, $S_{5\ldots J}$ is the sample standard deviation of tissues 5 to J and $\mu_{5\ldots J}^{max}$ is the maximum mean over tissue 5 to J.

3.3 Determining Bayesian Factor Threshold

Determining the Bayesian factor threshold is done in two steps using simulations in the first step and some adjustments based on our modifications in the second step.

3.3.1 Simulations for the original method

Simulations are performed for different types of tissue selectivity for the datasets in order to find thresholds. As mentioned in the (Van Deun, Hoijtink et al. 2009), choice of the Bayesian factor threshold depends on the number of tissues J, their number of replicates N_j for j=1…J and their effect sizes (Van Deun, Hoijtink et al. 2009) which are the parameters of simulations. The variance s, of a tissue, and overall mean expression level μ, in empirical data, are considered in order to generate the simulated data. The variance s of a tissue in the simulated data corresponds to the median standard deviation in the empirical data. The threshold is chosen such that there is a good balance between number of true and false positives which implies that the threshold is neither strict nor liberal. The two important factors of simulations are:

1. Amount of support in data for hypothesis H_1.
2. The effect size (Cohen 1969) which explains expression overlap between target tissue and rest of the tissues.

The amount of support in the data is divided into three levels:

1. All (J-1) tissues support the hypothesis H_1 completely meaning that mean of target tissue is greater than means of (J-1) tissues,
2. One tissue among (J-1) tissues has its mean either greater or equal to mean of target tissue meaning that the tissue doesn't support hypothesis H_1 and
3. All (J-1) tissues do not support the hypothesis H_1 meaning that mean of target tissue is less than means of (J-1) tissues.

The effect size δ is considered at two levels:

1. δ=0.5 and variance s of a tissue, reduced a little bit compared to original variance of a tissue in data, meaning target tissue sample will have considerable amount of overlap with other tissues sample and
2. δ=2 and variance s of a tissue, increased a little bit compared to original variance of a tissue in data, meaning the overlap between target tissue sample and rest of tissue samples is less.

The simulated data is generated from a normal distribution with variance s and mean μ for the target tissue, mean μ-δ for tissue supporting H_1 and mean μ+δ or mean μ for a tissue not supporting H_1. In the tissue-specific case only one tissue is compared to rest of tissues whereas in tissue-selective case minimum of tissue means is taken to compare to rest of the tissues. So in case of tissue-selective one of the target tissues will have mean μ+δ+ϵ and for the target tissue with minimum mean will have μ as its mean. The value ϵ depends on the number of target tissues in the hypothesis.

3.3.2 Adjustments for the modified method

After modifying the Bayesian factor method, to include the variability between tissues, the threshold is changed to a target tissue mean greater than rest of the tissue means at least by 50% of iterations which implies that $f_1 \geq 0.5$ (Section 3.1.1). If $f_1 \geq 0.5$ then $f_2 \leq 0.5$. Given conditions for f_1 and f_2 the Bayesian factor threshold is

$$BF \geq \frac{c_2}{c_1}.$$

Instead of using threshold for the Bayes factor we used a threshold directly for the posterior probability $p_g(\theta \in \Theta_1, \sigma^2 \,|y, H_1)$ for hypothesis H_1.

3.4 Vocabulary Mapping and Grouping

Datasets differ from each other by number of tissues, different tissues used, tissue sample size, type of species and different strains in species for some datasets. To compare the datasets despite of such differences vocabulary mapping and grouping is done.

3.4.1 Vocabulary Mapping

All five datasets have different tissue vocabularies but many tissues are shared among the datasets. In order to compare the datasets the tissues are grouped for all the five datasets based on tissues having similar function or name. The grouped tissues are shown in table listed in Appendix (Section A Vocabulary mapping table). Fetal and Cell line tissues are removed from all datasets as they are of no interest.

3.4.2 Grouping

The shared tissues among the datasets are grouped using the vocabulary mapping explained in (Section 3.4.1). In some datasets the sample size of tissues, which are to be grouped, differ which creates problem in computing the Bayesian factor and needs to be dealt with. This is shown in the below example. For a tissue:

Let T_1 be tissue 1 with sample – (x_1, x_2, x_3) and mean $X = (x_1+x_2+x_3)/3$,
T_2 is tissue 2 with sample - (y) and mean $Y = y$,
T_3 is tissue 3 with sample - $(z_1, z_2, z_3, z_4, z_5)$ and mean $Z = (z_1+z_2+z_3+z_4+z_5)/5$ and
T_4 is tissue 4 with sample - (u_1, u_2) and mean $U = (u_1+u_2)/2$.

The tissues T_1, T_2, T_3 and T_4 are similar tissues which are to be grouped to tissue T. To group the tissues T_1, T_2, T_3 and T_4 to tissue T the tissue means are compared and the tissue with highest mean is assigned to tissue T which creates the problem shown below for the above example.

For Gene 1: Let $X > Z > Y > U$ indicate that the sample of tissue T is T_1,
For Gene 2: Let $Y > Z > U > X$ indicate that the sample of tissue T is T_2,
For Gene 3: Let $Z > X > U > Y$ indicate that the sample of tissue T is T_3 and
For Gene 4: Let $U > Y > Z > X$ indicate that the sample of tissue T is T_4.

The sample size for the grouped tissue T changed for all four genes given that all four genes are from the same dataset. This change in sample of the grouped tissue T will make the computations harder as the sample size changes for every gene. In order to

avoid such cases the grouping algorithms is used. The two fundamental rules of the grouping algorithm are:

1. Minimum sample size that a grouped tissue can have is 2 and
2. Mean of grouped tissue should be maintained close to mean of original tissue as mean is really important in the Bayesian factor method.

The grouping algorithm is explained in the following steps:

1. Select the tissue with minimum sample size. If tissue with sample size = 1 exists then select the next minimum sample size so that final grouped tissue will have the minimum sample size which should be ≥ 2,

2. The grouped tissue sample size is same as minimum tissue sample size for each gene in the dataset.

3. For a tissue mean > mean of rest of tissues, which are to be grouped, and tissue sample size > minimum sample size then a subset of elements, length of subset = minimum sample size, are chosen from that tissue such that mean of the elements in subset is close to mean of that tissue. The subset becomes the sample of the grouped tissue.

4. For a tissue mean > mean of rest of tissues, which are to be grouped, and tissue sample size = 1 then the element in that tissue is repeated as per the minimum sample size such that the mean is conserved. Also the repeated elements are added with mean of standard deviation of the tissues which are to be grouped. The set of repeated elements added with mean of standard deviation becomes the sample of the grouped tissue.

5. Repeat the step 3 and 4 for all the genes in the dataset.

The steps in grouping algorithm are illustrated for the above example:

Among the four tissues the tissue with minimum sample size is T_2. But step 1 of the grouping algorithm says that minimum sample size should be ≥ 2 which is satisfied by tissue T_4. The grouped tissue T will have sample size = 2.

1. For Gene 1 two elements from tissue T_1 are selected such that mean of two selected elements is $\approx X$.
2. For Gene 2 the sample of tissue T is (y + mean(SD_1, SD_3, SD_4), y - mean(SD_1, SD_3, SD_4)).
3. For Gene 3 two elements from tissue T_3 are selected such that mean of two selected elements is $\approx Z$.
4. For Gene 4 the grouped tissue T is T_4 as T_4 has sample size = 2.

Where SD_i is standard deviation of tissue i=1, 3 and 4. The concept of grouping ensures that grouped tissues have the same sample size in every gene of the dataset which simplifies the computations. The minimum sample size that a grouped tissue can have is

2 because the standard deviation needs to be defined in the calculation of the Bayesian factors.

3.5 Variability between Strains and within Strains

Due to different strains used for the different tissues in Rat an analysis of variance between and within the Rat strains is done to investigate whether it is possible to combine replicates from the different strains. The hypotheses to be tested are

$$H_0 : \mu_{combined} = \mu_{strain1} = \mu_{strain2} = \cdots = \mu_{strainN} ,$$
$$H_1 : \text{not } H_0 ,$$

where *N* represents number of strains in the dataset. The hypothesis testing is done by following methods:

1. ANOVA and,
2. Variance Components Model II ANOVA

The details of the methods are listed in Appendix (Section D-1and D-2).

3.6 Mapping Genes to Probesets

In the datasets number of probesets used is greater than amount of genes covered. The annotation file of a dataset has the information between probesets and genes. The information regarding the probesets and number of genes covered by the dataset is shown in Table 1.

Table 1 Amount of probesets used and genes covered in datasets

DATASETS	Total Genes	Total Probesets	Mapped Probesets	Annotation File
Human(GDS3113)	16367	32878	18002	GPL2986.annot
Human(GDS596)	13696	22283	21169	GPL96.annot
Human(GSE7307)	21076	54675	42369	GPL570.annot
Mouse(GSE9954)	23308	45101	39331	GPL1261.annot
Rat(GSE952)	5349	8799	7722	GPL85.annot

The total genes column mentioned in the Table 1 does not include duplicates. Every dataset has its own annotation file which gives detailed description of relation between genes and probesets. Utilization of annotation files to capture the relation is implemented in an R script for all datasets. The concept of regular expression is used in R script. Sample of annotation file of a dataset is shown in Table 2.

Table 2 Sample of GDS3113 dataset annotation file showing only first three columns

Probe ID	Gene title	Gene Symbol
156427	-------	-------
131316	Gephyrin	GPHN

128500	Small EDRK-rich factor 1B/// Small EDRK-rich factor 1A	SERF1B///SERF1A

There are many columns in the annotation file but only the first and third columns are useful and the rest of the columns are ignored. In the R script the mapping is done between probeset ID and gene symbol. Only the probesets that are mapped to genes are considered for the Bayesian factor method.

3.7 Implementation

The whole project is done in the R software (Team 2008). For the Bayesian factor method we have used the BayesianIUT package (Van Deun, Hoijtink et al. 2009) and modified the R script for the modified Bayesian factor. All the figures are also generated in R. The regular expression used for removing duplicates, while mapping probesets to genes, is also done in R.

4 Results

Before optimizing the parameters for the modified Bayesian factor method (Section 3.2) we have checked whether the Rat (GSE952) data can be combined even though it has strains. We have done ANOVA (Section D-1) and Variance Components (Section D-2) for five tissues that are common in the three strains of the Rat (GSE952) dataset. Interestingly we found that the strains cannot be combined as the ratio of variance between and within the strains is quite high and also the F value < 0.05 significant probesets are really less compared to probesets with insignificant F value, out of 8769 probesets, in the dataset. The results are shown in Table 3.

Table 3 Results of ANOVA and Variance Components of Five tissue in the Rat (GSE952) dataset

Tissue	Sample Size in Wistar	Sample Size in Wistar Ky	Sample Size in Sprague	$F < 0.05$	$\frac{\overline{\sigma^2}}{\overline{\omega^2}}$
Frontal Cortex	10	6	6	13.1%	1.283739
Cerebellum	6	5	6	11.75%	1.162023
Cerebral Cortex	3	4	6	11.7%	1.205392
Amygdala	6	1	3	16%	1.883162
Ventral Striatum	3	2	3	4.8%	0.8760915

We have plotted the histograms of logarithm of variance within and between strains for five tissues for the method ANOVA and the plots are listed in Appendix (Section B The Histogram Plots). As the strains cannot be combined we have selected the strain Sprague Dawley in the Rat (GSE952) dataset which covers 26 tissues which are grouped to 12 tissues using vocabulary mapping (Section 3.4.1). While grouping the tissues in the Rat (GSE952) dataset we found some tissue samples that do not have any replicates so we used step 4 of the grouping algorithm (Section 3.4.2). The reason for taking the mean of the standard deviation is based on the training data and the density plots of standard deviation for the Rat (GSE952) dataset. To see how the standard deviation varies for a tissue in the Rat (GSE952) dataset we have divided the Rat (GSE952) data into four groups based on mean of that tissue per gene which is as follows:

Group 1 (G1): $12 < \mu \leq 16$,
Group 2 (G2): $8 < \mu \leq 12$,
Group 3 (G3): $4 < \mu \leq 8$ and
Group 4 (G4): $\mu \leq 4$.

We have done this for tissues mentioned in Table 3 and the density plots of the tissues are listed in Appendix (Section C The Density Plots). All the density plots (Section C The Density Plots) follow similar pattern which is higher the mean of the tissue lower its standard deviation. So we have decided to take mean of standard deviation for step 4 in grouping (Section 3.4.2). After fixing the Rat (GSE952) dataset we moved on to optimizing the parameters C and posterior probability threshold for all five datasets for different gene expression patterns. The choice of the parameter C for a dataset is completely based on training data of that dataset. But for 3-selective and 4-selective cases the choice of parameter C is based on the choice of either tissue-specific case or 2-selective case because of coverage of different genes by training data (Section 2.2). We

have optimized the parameters for all datasets using the training data and the results of optimized parameters for all five datasets for different gene expression patterns are shown in Table 4.

Table 4 Optimized C value and posterior probability threshold (PPT) of H_1 for five datasets

Datasets	Specific Case		2-Selective		3-Selective		4-Selective	
	C	PPT	C	PPT	C	PPT	C	PPT
Human(GDS3113)	1.79	55%	1.92	77%	1.79	50%	1.79	50%
Human(GDS596)	1.64	50%	1.64	50%	1.79	50%	1.79	50%
Human(GSE7307)	1.79	50%	1.64	70%	1.79	70%	1.79	68%
Mouse(GSE9954)	1.79	64%	3.5	60%	3.5	80%	3.5	80%
Rat(GSE952)	1.6	68%	2	70%	2	75%	2	70%

As explained in (Section 3.3.2) for posterior probability threshold the target tissue mean should be greater than rest of the tissues at least by 50% of iterations but we considered it to be more than 70% due to training data except for Human (GDS596) dataset which is exactly 50% can be seen in Table 4. The variation in the posterior probability threshold for a dataset depends not only on the training data but also on the number of tissues and simulations (Section 3.3.1) for respective datasets. The variation in C value for a dataset is completely based on the training data. The optimized parameters are not that bad but on the contrary they are not that good either due to lack of training data genes for 3-selective and 4-selective cases. We still need more training data for better optimized results. Using the optimized parameters we ran the modified Bayesian factor method (Section 3.2) for each probeset in all datasets for 10,000 iterations and the results are shown in Table 5.

Table 5 Result of Bayesian Factor method on all datasets in terms of probesets

Datasets	Specific	2-selective	3-selective	4-selective	Total
Human(GDS3113)	2317	540	307	170	3334
Human(GDS596)	580	118	8	6	712
Human(GSE7307)	2482	335	55	17	2889
Mouse(GSE9954)	7958	1007	444	147	9556
Rat(GSE952)	777	384	131	79	1371

The number of genes identified in the Human (GDS596) dataset is less compared to the total number of genes that the dataset covers. The results in Table 5 indicate that there are 2317 probesets that are tissue-specific but does not give any details of number of the genes that are tissue-specific. So we have mapped the probesets to genes (Section 3.6) and removed duplicates. In the process of removing duplicates two cases are considered where a gene can have many probesets and a probeset can indicate to set of genes:

Gene ----> Probe
HFE ---> 207283
HFE ---> 226876
GGTLC2///GGT3P///GGT1 ---> 199042

We have considered only the mapped probesets because we wanted to compare the Bayesian factor method results among the datasets. It would be problematic if unmapped probesets are also considered for comparison among the datasets. After mapping probesets to genes, the results of Bayesian factor in terms of number of genes identified are shown in Table 6.

Table 6 Result of Bayesian Factor method on all datasets in terms of genes

Datasets	Specific	2-selective	3-selective	4-selective	Total
Human(GDS3113)	2275	539	307	170	3291
Human(GDS596)	554	104	11	6	675
Human(GSE7307)	2168	287	56	18	2529
Mouse(GSE9954)	6911	895	402	151	8359
Rat(GSE952)	709	356	133	77	1275

The results in Table 6 do not include duplicates. The Human (GDS596) dataset has much noise in the data resulting in less number of genes detected compared to rest of the datasets. For the Human (GDS596) dataset we cannot lower down the parameters anymore as they are the minimum that we can consider. For tissue-specific case the number of genes detected is high due to the fact that training data for all datasets covers tissue-specific genes mostly compared to other tissue-selective genes. So the number of genes detected decreases from left to right in a row in Table 6. Also for Rat (GSE952) dataset the number of genes detected is less because we have considered data only from the strain Sprague Dawley there by losing data from other strains. After mapping probesets to genes and getting the Bayesian factor results in terms of genes we were interested in finding the overlap in tissue-specific genes across all five datasets. The percent of overlap between two datasets is taken as

$$\%\ of\ overlap = \frac{n(similar)}{n(A) + n(B) - n(similar)}\ ,$$

where $n(A)$ is number of genes in dataset A, $n(B)$ is number of genes in dataset B and $n(similar)$ is number of genes present in both dataset A and dataset B. In some datasets the gene symbols are in lower case and in some datasets they are in upper case. While trying to calculate the number of genes that are present in both datasets we used case insensitive comparison to avoid such cases. The result of tissue-specific genes overlap across all five datasets is shown in Table 7.

Table 7 Overlap of tissue-specific genes across all five datasets

DATASETS	GDS3113	GDS596	GSE7307	GSE9954	GSE952
Human(GDS3113)	2275	207	727	835	120
Human(GDS596)	7.895	554	259	188	64
Human(GSE7307)	19.56	10.52	2168	672	105
Mouse(GSE9954)	9.99	2.583	7.99	6911	229
Rat(GSE952)	4.19	5.34	3.79	3.1	709

The diagonal in Table 7 represents number of tissue-specific genes in the respective datasets. The lower triangle of Table 7 represents the percent of overlap between the datasets and the upper triangle represents the number of similar genes between datasets. The overlap among the three Human datasets (GDS3113, GDS596, and GSE7307) is high. For the Mouse (GSE952) dataset overlap with the Human (GDS3113) and Human (GDS7307) datasets is high compared to overlap with the Human (GDS596) dataset and this difference in overlap is due to less number of genes identified for the Human (GDS596) dataset. We have also done the overlap of 2-selective, 3-selective and 4-selective genes among the datasets and the result of the overlap are shown in Table 8-Table 10.

Table 8 Overlap of 2-selective genes across all five datasets

DATASETS	GDS3113	GDS596	GSE7307	GSE9954	GSE952
Human(GDS3113)	539	9	41	53	25
Human(GDS596)	1.42	104	24	11	13
Human(GSE7307)	5.22	6.54	287	21	10
Mouse(GSE9954)	3.84	1.113	1.81	895	34
Rat(GSE952)	2.874	2.91	1.58	2.794	356

Table 9 Overlap of 3-selective genes across all five datasets

DATASETS	GDS3113	GDS596	GSE7307	GSE9954	GSE952
Human(GDS3113)	307	0	10	31	5
Human(GDS596)	0	11	4	0	1
Human(GSE7307)	2.83	6.35	56	6	1
Mouse(GSE9954)	4.57	0	1.33	402	13
Rat(GSE952)	1.15	0.7	0.532	2.5	133

Table 10 Overlap of 4-selective genes across all five datasets

DATASETS	GDS3113	GDS596	GSE7307	GSE9954	GSE952
Human(GDS3113)	170	0	3	3	0
Human(GDS596)	0	6	0	2	2
Human(GSE7307)	1.62	0	18	0	0
Mouse(GSE9954)	0.94	1.29	0	151	3
Rat(GSE952)	0	2.47	0	1.34	77

For 2-selective, 3-selective and 4-selective cases the overlap between the Human datasets decreases which can be observed in the above tables. This decrease in overlap was due to lack of sufficient tissue-selective genes in training data.

5 Discussion and Conclusion

The main aim of this project was to extend and modify the Bayesian factor method so that it can identify different gene expression patterns. As mentioned in the Results (Section 4) that optimized parameters are not bad but they are not good at the same time. This can be improved if training data can cover a higher number of different tissue-selective genes. The number of genes identified by the Bayesian factor method on the Human (GDS596) dataset is really small which implies that the data has a lot of noise. Having less noisy data would increase the chance of detecting a tissue-selective gene. The amount of tissues covered varies among the datasets and there are cases where some tissues are present in some datasets while absent in others. This would complicate the comparison of results between datasets. It would be better if all datasets cover the same set of tissues.

The datasets does not account for the variability between tissues which has a huge impact on the results giving the number of false positives. In the Rat (GSE952) dataset different strains are used in the experiment which we showed cannot be treated as replicates. The quality of the data depends on the way the experiment is carried out. In recent years new upcoming sequencing technologies have emerged, such as RNA-Seq (Wang, Gerstein et al. 2009). The RNA-Seq technology can provide better quality transciptome data where the Bayesian factor method can be a powerful tool.

Another problem is that training data for most of the species lack tissue-selective genes. In the original implementation of the Bayesian factor method the parameters were chosen from simulations alone. After our modification the parameters settings depend on training data. Higher quantity of training data will lead to better performance. There were not enough 2-selective, 3-selective and 4-selective genes in training data for the Mouse (GSE9954) and Rat (GSE952) species. In the future we expect more tissue-selective genes to be present resulting in better performance of the modified Bayesian factor method.

Despite of the above mentioned drawbacks, the performance of the modified Bayesian factor method was quite satisfactory as it was able to correctly predict many tissue-selective genes. It is also possible to build better models depicting Gene-Tissue relationships more accurately using the Bayesian framework - which may increase the chance of identifying suitable biomarkers and drug targets.

References

Chib, S. (1995). "Marginal likelihood from Gibbs output." Journal of the American Statistical Association **90**: 1313-1321.

Cohen, J. (1969). Statistical Power Analysis for the Behavioral Sciences, Academic Press, New York.

Dezso, Z., Y. Nikolsky, et al. (2008). "A comprehensive functional analysis of tissue specificity of human gene expression." BMC Biol **6**: 49.

Freilich, S., T. Massingham, et al. (2005). "Relationship between the tissue-specificity of mouse gene expression and the evolutionary origin and function of the proteins." Genome Biol **6**(7): R56.

Hu, E., Z. Chen, et al. (2001). "Rapid isolation of tissue-specific genes from rat kidney." Exp Nephrol **9**(2): 156-164.

Irene Klugkist, H. H. (2007). "The Bayes factor for inequality and about equality constrained models." Computational Statistics & Data Analysis **51**: 6367– 6379.

Kadota, K., J. Ye, et al. (2006). "ROKU: a novel method for identification of tissue-specific genes." BMC Bioinformatics **7**: 294.

Knoll, K. E., J. L. Pietrusz, et al. (2005). "Tissue-specific transcriptome responses in rats with early streptozotocin-induced diabetes." Physiol Genomics **21**(2): 222-229.

Kouadjo, K. E., Y. Nishida, et al. (2007). "Housekeeping and tissue-specific genes in mouse tissues." BMC Genomics **8**: 127.

Liang, S., Y. Li, et al. (2006). "Detecting and profiling tissue-selective genes." Physiol Genomics **26**(2): 158-162.

Lieven Thorrez, K. V. D., Léon-Charles Tranchevent, Leentje Van Lommel, Kristof Engelen, Kathleen Marchal, Yves Moreau, Iven Van Mechelen, Frans Schuit (2008). "Using Ribosomal Protein Genes as Reference: A Tale of Caution." PLoS One **3**(3): e1854.

Silver, N., E. Cotroneo, et al. (2008). "Selection of housekeeping genes for gene expression studies in the adult rat submandibular gland under normal, inflamed, atrophic and regenerative states." BMC Mol Biol **9**: 64.

Stansberg, C., A. O. Vik-Mo, et al. (2007). "Gene expression profiles in rat brain disclose CNS signature genes and regional patterns of functional specialisation." BMC Genomics **8**: 94.

Su, A. I., T. Wiltshire, et al. (2004). "A gene atlas of the mouse and human protein-encoding transcriptomes." Proc Natl Acad Sci U S A **101**(16): 6062-6067.

Team, R. D. C. (2008). R: A language and environment for statistical computing, reference index version 2.13.1. Vienna, Austria, R Foundation for Statistical Computing.

Walker, J. R., A. I. Su, et al. (2004). "Applications of a rat multiple tissue gene expression data set." Genome Res **14**(4): 742-749.

Van Deun, K., H. Hoijtink, et al. (2009). "Testing the hypothesis of tissue selectivity: the intersection-union test and a Bayesian approach." Bioinformatics **25**(19): 2588-2594.

Wang, Z., M. Gerstein, et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nat Rev Genet **10**(1): 57-63.

Xiao, S. J., C. Zhang, et al. (2010). "TiSGeD: a database for tissue-specific genes." Bioinformatics **26**(9): 1273-1275.

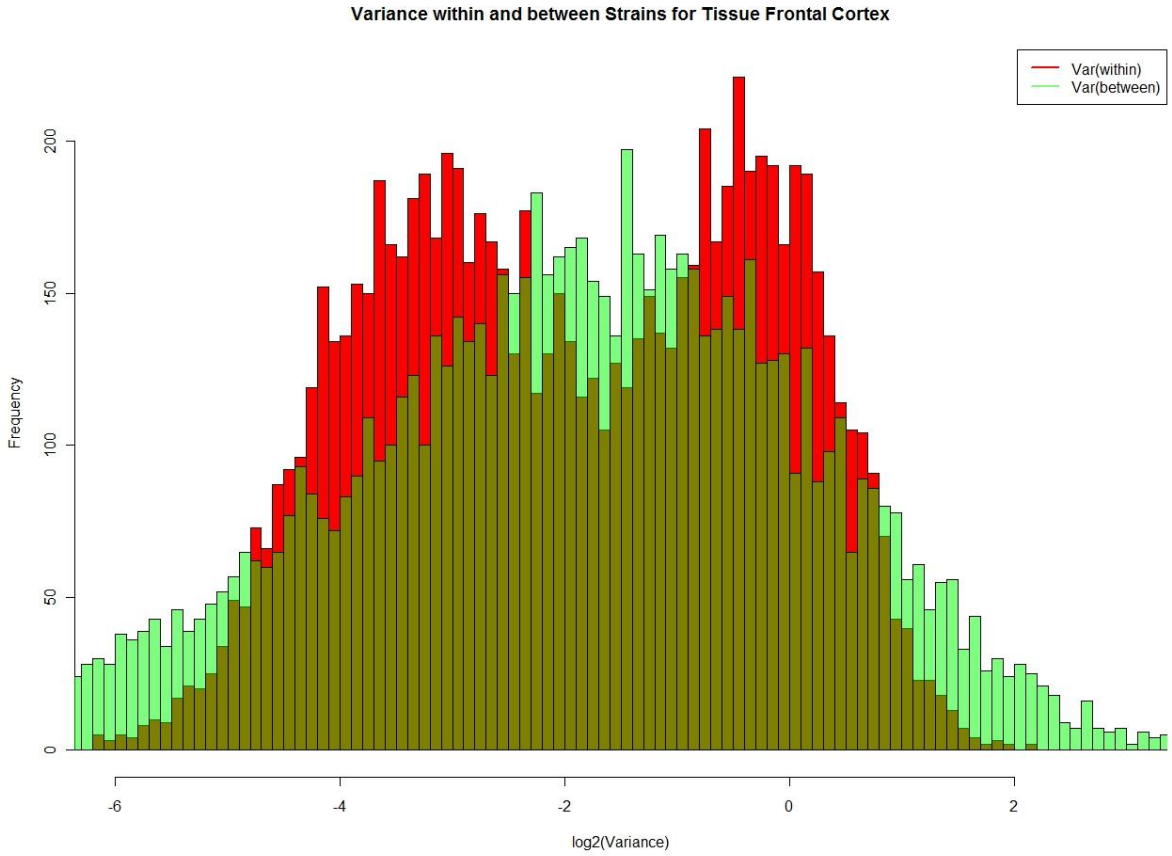
Appendix

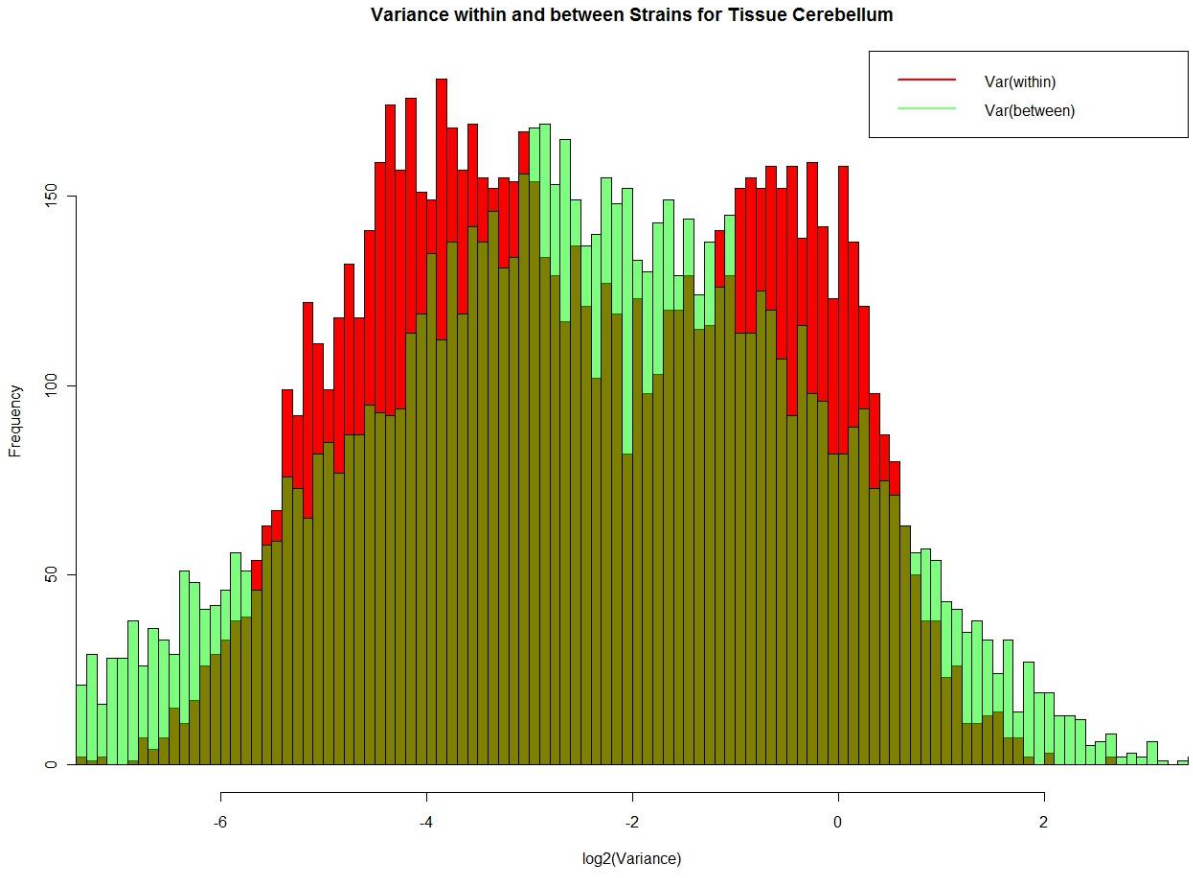
A Vocabulary mapping table

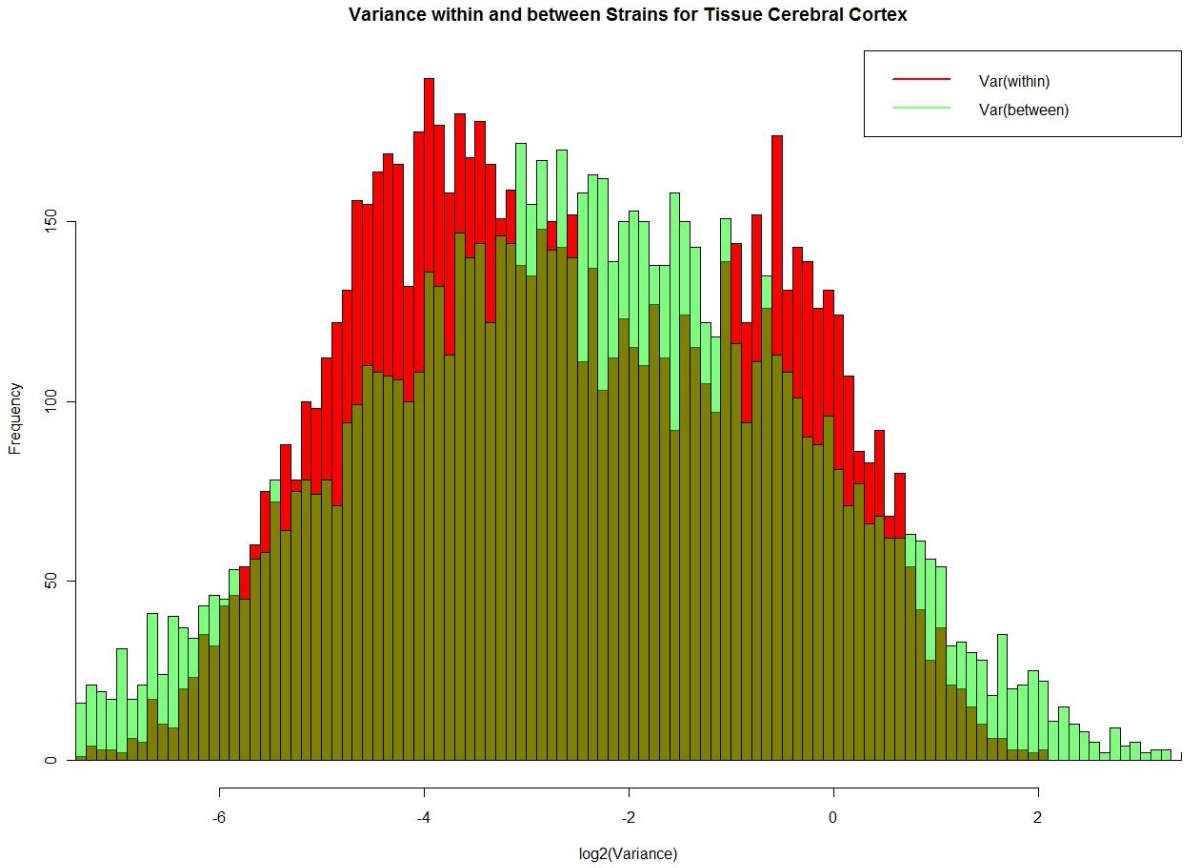
Grouped	Original	Grouped	Original
Adipocyte	Adipose Tissue	Large Intestine	Large Intestine
	Adipose Tissue Omental	Leukemia	Leukemia Chronic Myelogenous K-562
	Adipose Tissue Subcutaneous		Leukemia Promyelocytic-HL-60
Adrenal	Adrenal Gland		Leukemia lymphoblastic(MOLT-4)
	Adrenal Gland Cortex	Liver	Liver
Appendix	Appendix	Lung	Lung
Bone Marrow	Bone Marrow	Lymph	721_B_lymphoblasts
Breast	Breast		Lymph node
Bronchus	Bronchus		Lymphoma Burkitts(Daudi)
Cerebellum	Cerebellar Hemisphere		Lymphoma Burkitts(Raji)
	Cerebellar Vermis	MammaryGland	Mammary Gland
	Cerebellum	Muscle	Skeletal Muscle
	Cerebellum Peduncles		Smooth Muscle
Cervix	Cervix		Cardiac Myocytes
CNS	Accumbens		Skeletal Muscle Superior Quadracep
	Amygdala	Myometrium	Myometrium
	Brain	Nerve	Ciliary Ganglion
	Caudate Nucleus		Dorsal Root Ganglion
	Cerebral Cortex		Superior Cervical Ganglion
	Cingulate Cortex		Trigeminal Ganglion
	Corpus Callosum	Ovary	Ovary
	Dorsal Raphe	Pancreas	Pancreas
	Dorsal Striatum		Pancreatic Islet
	Frontal Cortex	Pineal	Pineal
	Globus Pallidus	Placenta	Placenta
	Gloubus Pallidum External	Prostate	Prostate
	Gloubus Pallidum Internal		Prostate Gland
	Hippocampus	Salivary Gland	Salivary Gland
	Hypothalamus	Seminal Vesicle	Seminal Vesicle
	Locus Coeruleus	Skin	Skin
	Medulla Oblongata	Small Intestine	Small Intestine BD
	Nucleus Accumbens Core		Small Intestine Duodenum

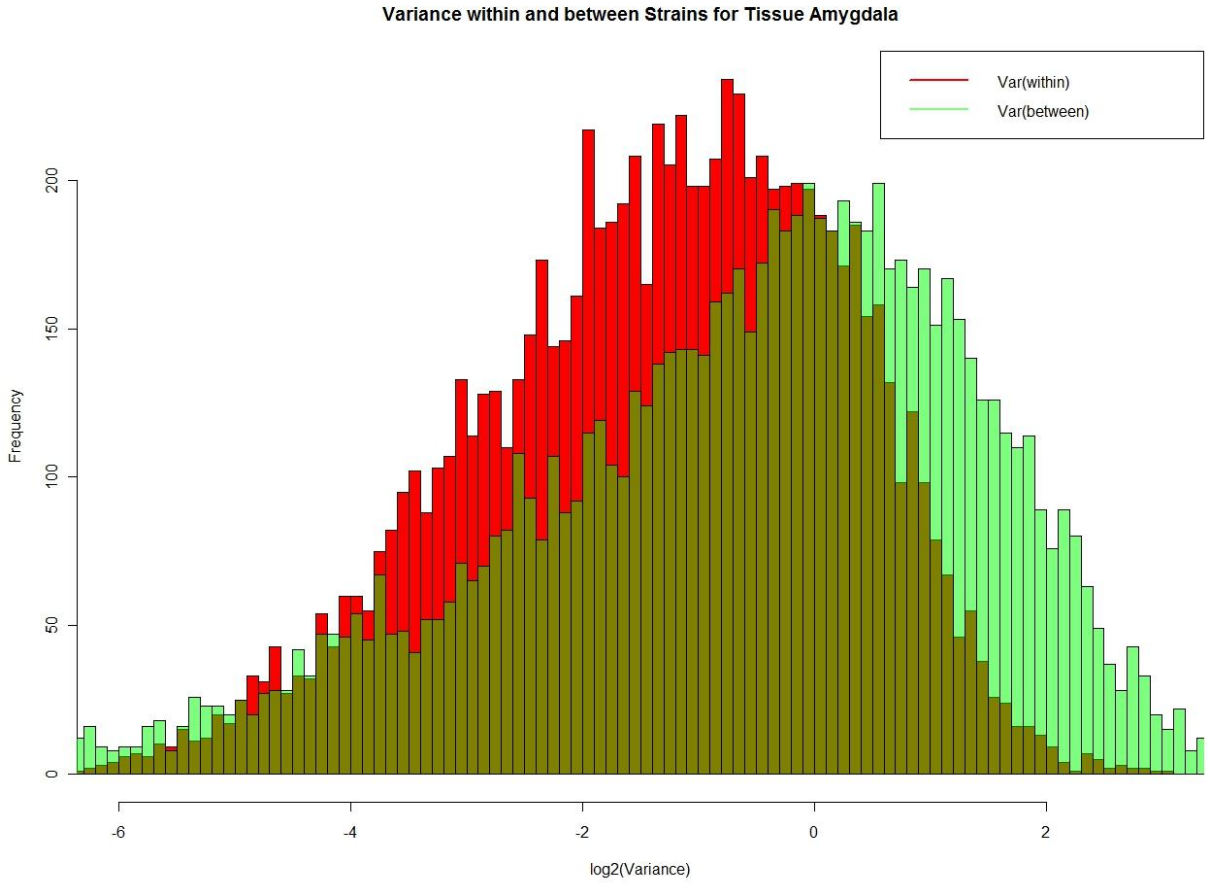
	Nucleus Accumbens shell		Small Intestine Ileum
	Occipital Lobe		Small Intestine Jejunum
	Olfactory Bulb	Spinal Cord	Spinal Cord
	Parietal Lobe	Spleen	Spleen
	Pituitary	Stomach	Stomach
	Pons		Stomach Cardiac
	Prefrontal Cortex		Stomach Fundus
	Primary Cortical Neurons		Stomach Pyloric
	Putamen	Testis	Testis
	Substantia Nigra		Testis Germ Cell
	Substantia Nigra Pars Compacta		Testis Intersitial
	Substantia Nigra Reticulata		Testis Leydig Cell
	Subthalamic Nucleus		Testis Seminiferous Tubule
	Temporal Lobe	Thymus	Thymus Gland
	Thalamus	Thyroid	Thyroid Gland
	Ventral Striatum	Tongue	Tongue
	Ventral Tegmental Area		Tongue Main Corpus
	Vestibular Nuclei Superior		Tongue Superior With Papillae
Colon	Colon	Tonsil	Tonsil
	Colon Cecum	Trachea	Trachea
Diaphragm	Diaphragm	Urethra	Urethra
Endometrium	Endometrium	Uterus	Uterus
Esophagus	Esophagus		Uterus Corpus
Eye	Eye	Vagina	Vagina
	Cornea	Vessel	Aorta
Fallopian Tube	Fallopian Tube		Coronary Artery
Heart	Atrioventricular Node		Saphenous Vein
	Heart		Vena Cava
	Heart Atrium	Vulva	Vulva
	Heart Ventricle		
Kidney	Kidney		
	Kidney Cortex		
	Kidney Medulla		

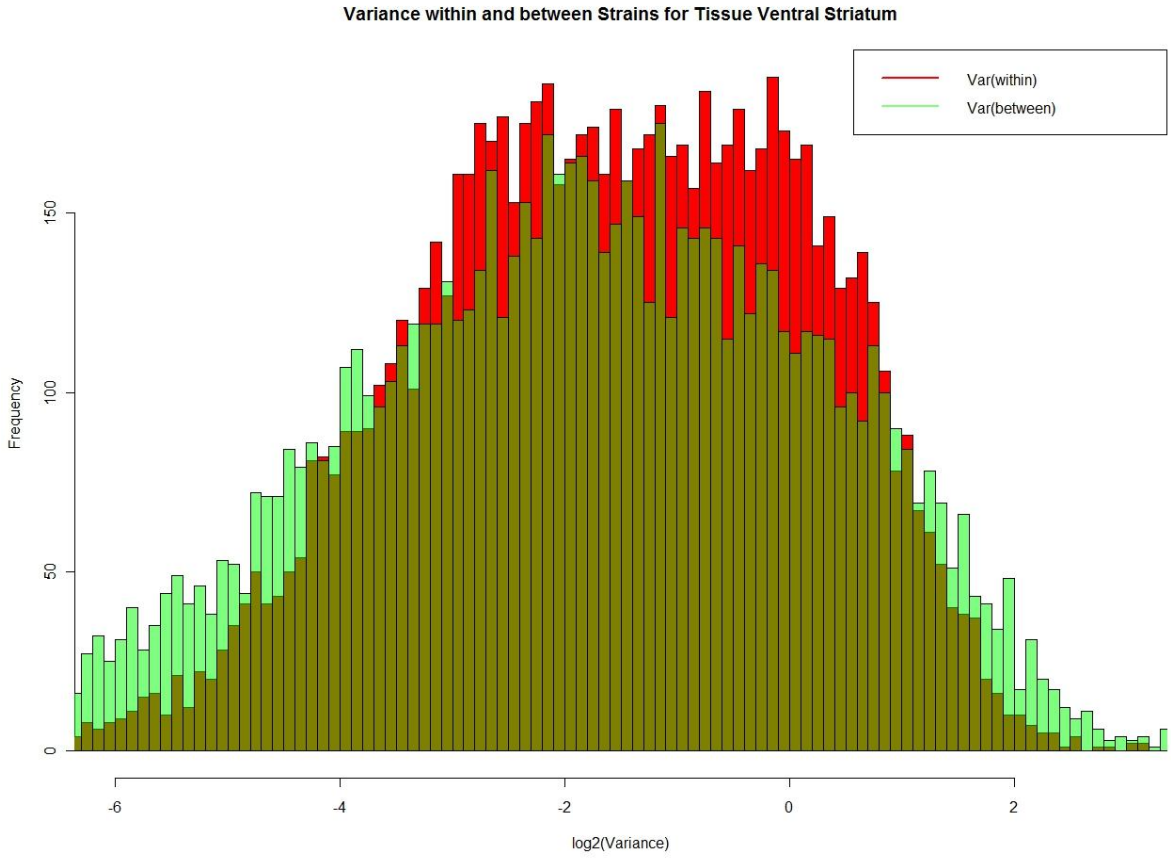
B The Histogram Plots






Variance within and between Strains for Tissue Cerebral Cortex
Var(within)
Var(between)
Frequency
0
50
100
150
-6
-4
-2
0
2
log2(Variance)

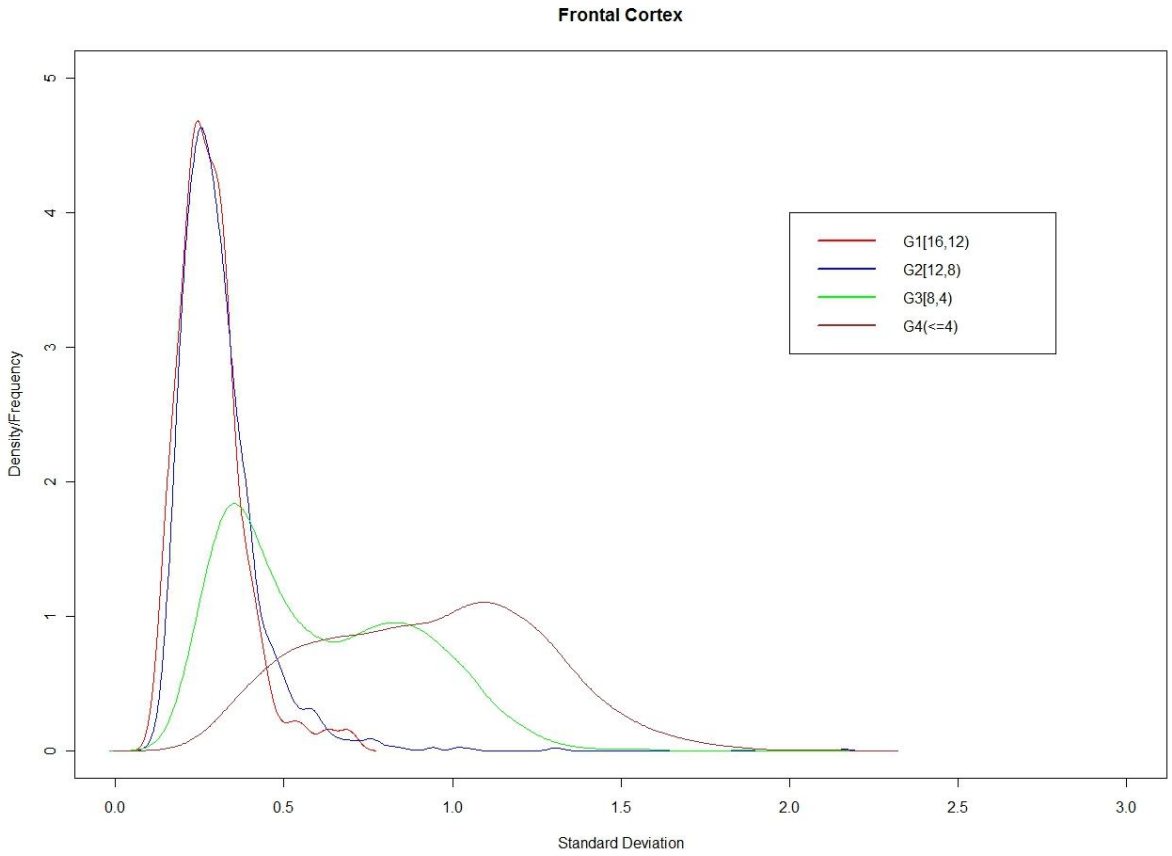

Variance within and between Strains for Tissue Amygdala
Var(within)
Var(between)
Frequency
0
50
100
150
200
-6
-4
-2
0
2
log2(Variance)

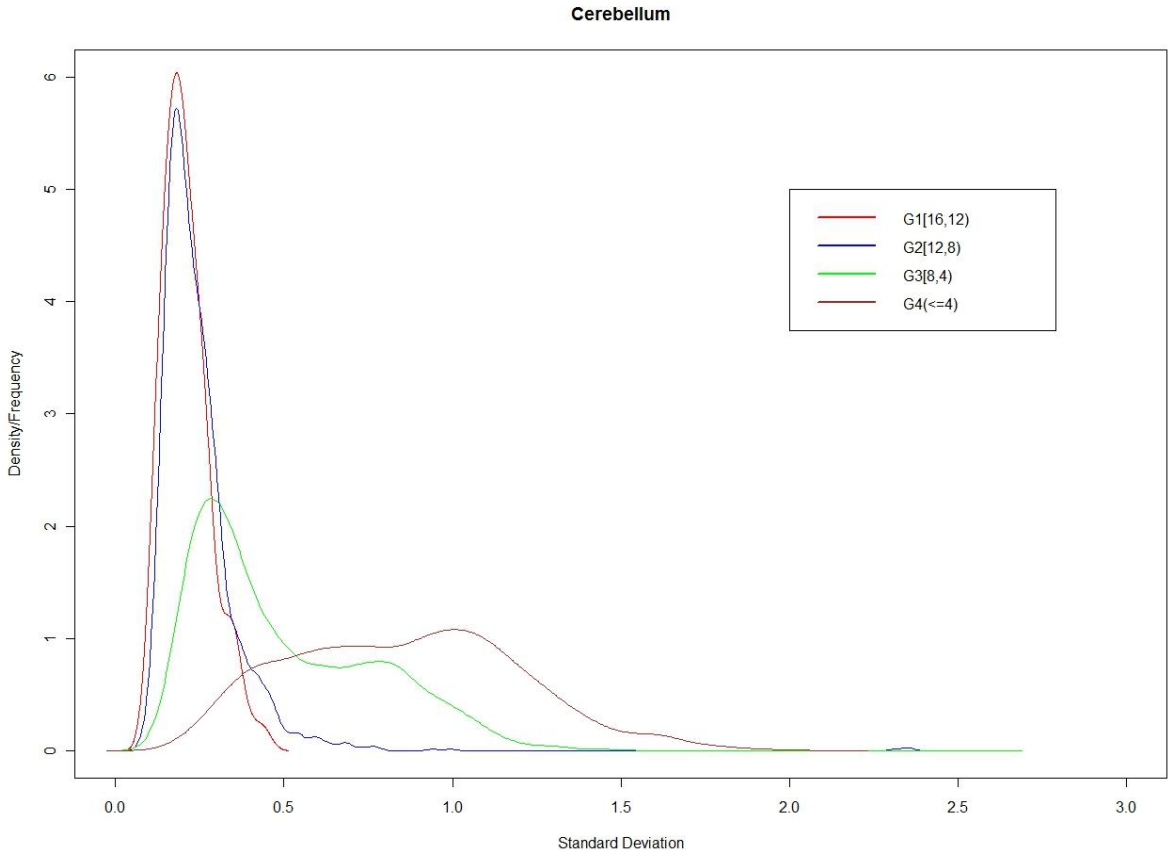

Variance within and between Strains for Tissue Ventral Striatum
Var(within)
Var(between)
Frequency
0
50
100
150
-6
-4
-2
0
2
log2(Variance)

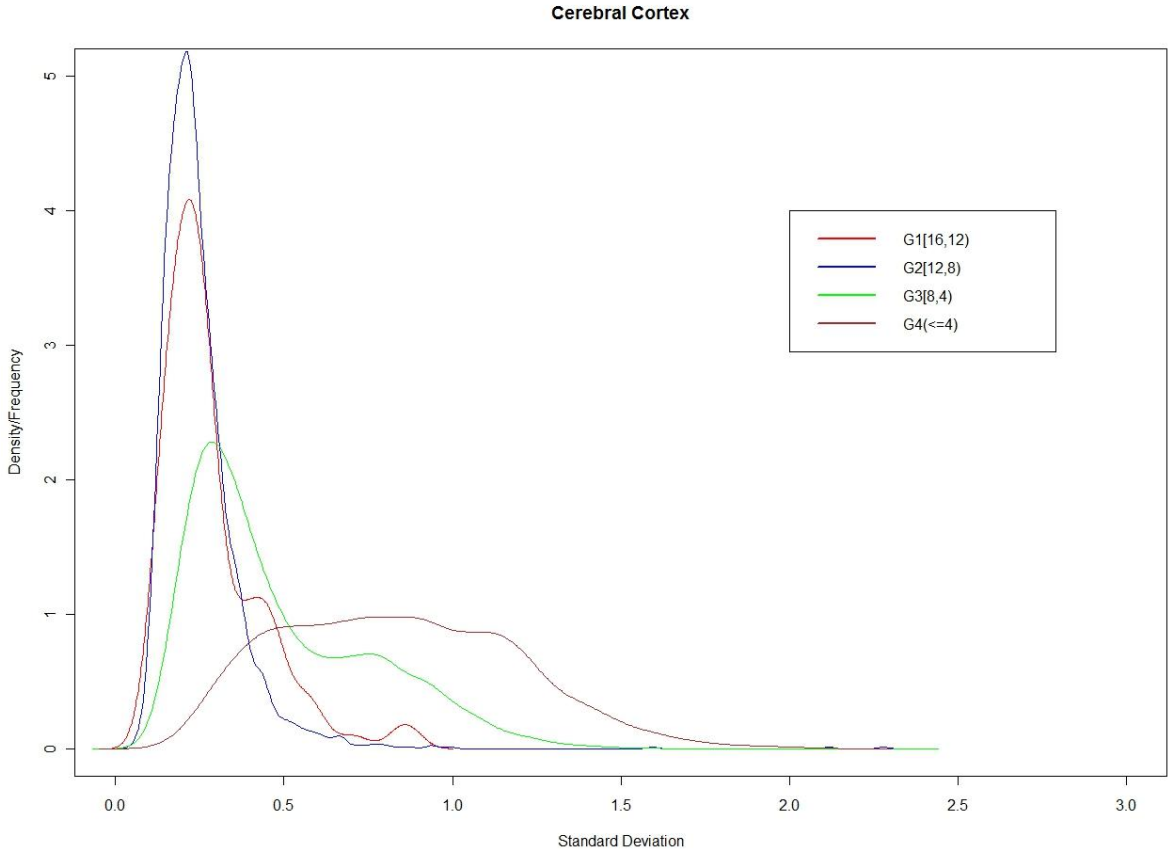
C The Density Plots

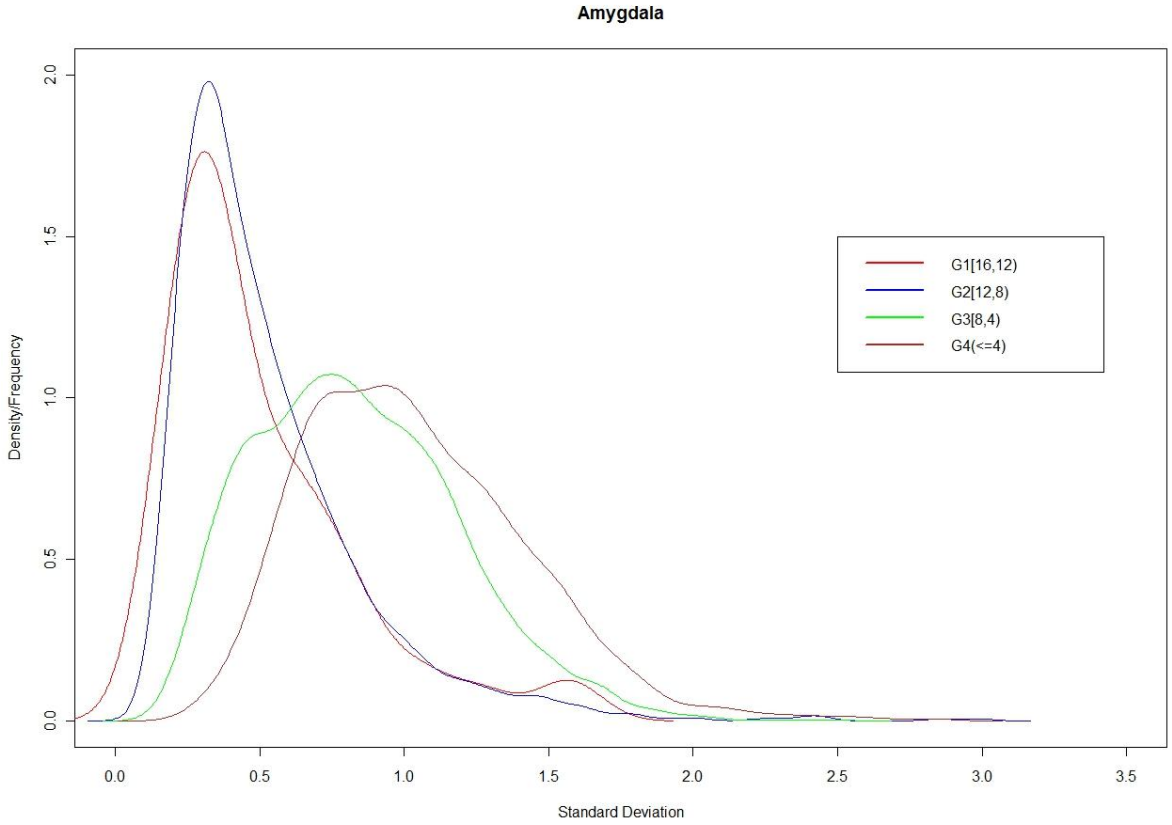
The density plots show the variation of the standard deviation for a tissue in the Rat (GSE952) dataset. The Rat (GSE952) data is divided into four groups based on mean of the tissue per gene in the dataset. The groups are:

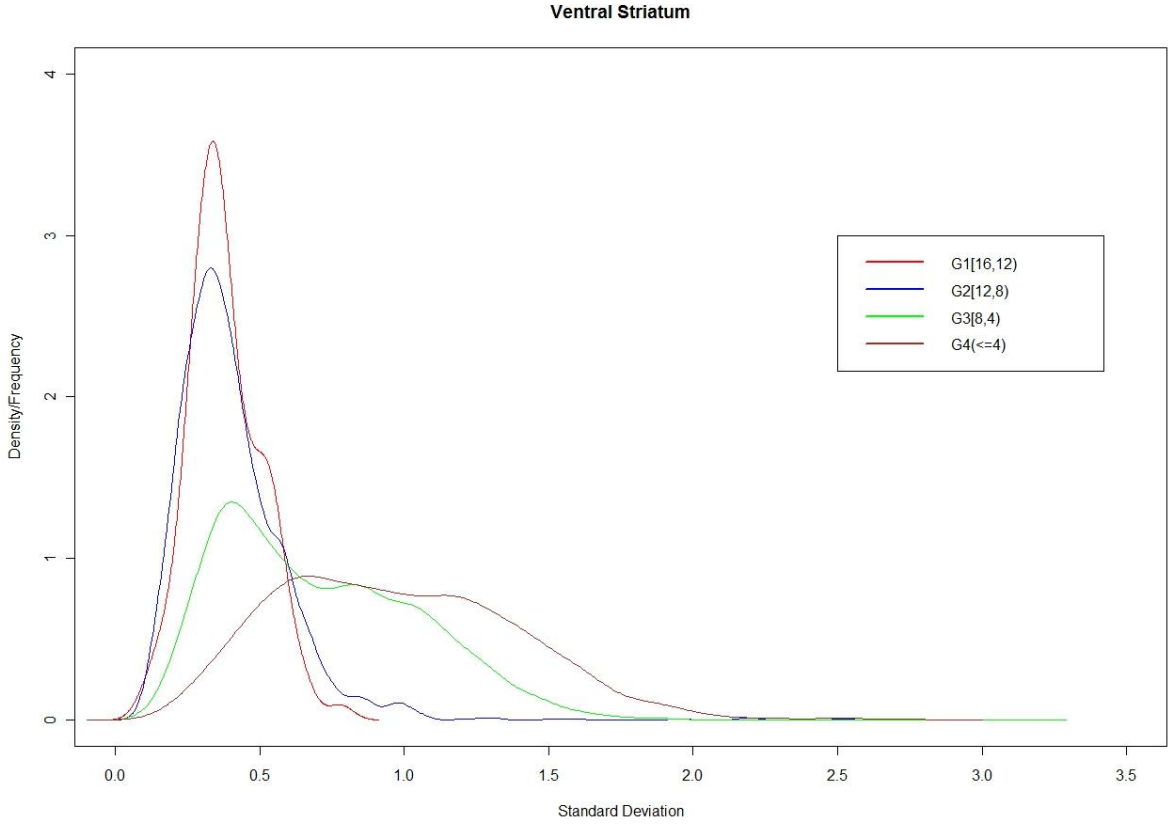
Group 1 (G1): $12 < \mu \leq 16$,
Group 2 (G2): $8 < \mu \leq 12$,
Group 3 (G3): $4 < \mu \leq 8$ and
Group 4 (G4): $\mu \leq 4$.




Cerebellum
Density/Frequency
Standard Deviation
G1[16,12)
G2[12,8)
G3[8,4)
G4(<=4)


Cerebral Cortex
Density/Frequency
Standard Deviation
G1[16,12)
G2[12,8)
G3[8,4)
G4(<=4)


Amygdala
Density/Frequency
Standard Deviation
G1[16,12)
G2[12,8)
G3[8,4)
G4(<=4)


Ventral Striatum
Density/Frequency
Standard Deviation
G1[16,12)
G2[12,8)
G3[8,4)
G4(<=4)

D-1 ANOVA

The variance between and within the strains is calculated for each gene. The number of strains in the dataset is *N*. For a tissue:

Let n_i be sample size of strain *i* where *i*=1… *N*.

$\bar{X}$ is mean of the tissue and $\bar{X}_i$ is mean of strain *i*=1… *N*.

S_i^2 is the variance of the strain *i*=1… *N*.

$$SS_B = \sum_{i=1}^{N} n_i (\bar{X}_i - \bar{X})^2 \, , df_B = N - 1 \, ,$$

$$S_B^2 = \frac{SS_B}{df_B} \, ,$$

$$S_W^2 = \frac{\sum_{i=1}^{N} (n_i - 1) S_i^2}{\sum_{i=1}^{N} n_i - 1} \, ,$$

where S_B^2 is variance between strains, S_W^2 is variance within strains, SS_B is sum of squares between strains and df_B is degrees of freedom for variance between strains. *F* value is calculated for the ratio of variance between and within strains.

$$F_{stat} = \frac{S_B^2}{S_W^2} \, ,$$

where F_{stat} is calculated using *F* distribution table. The *F* value < 0.05 is considered significant.

D-2 Variance Components

The variance between and within strains are calculated bit differently from ANOVA in (Section D-1). For a tissue in a probeset:

Let n_i be sample size of strain *i* where *i*=1… *N*.

SS_B is sum of squares between strains; SS_W is sum of squares within strains,

$$E\left[\frac{SS_B}{(N-1)}\right] = \sigma^2 + \frac{1}{n}\omega^2 \, ,$$

$$E\left[\frac{SS_W}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_N - 1)}\right] = \omega^2 ,$$

$$E\left[\frac{SS_B}{(N-1)} - \frac{1}{n}\frac{SS_W}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_N - 1)}\right] = \sigma^2 \, ,$$

where ω^2 is variance within strains also pooled variance, σ^2 is variance between the strains. Taking the average over all the probesets in the dataset will give $\overline{\omega^2}$ and $\overline{\sigma^2}$. If the ratio $\frac{\overline{\sigma^2}}{\overline{\omega^2}}$ is high then the strains cannot be mixed.