CHALMERS

Tissue-Specific Gene Expression Analysis

A large-scale meta-analysis of gene expression pattern using microarray data

Master of Science Thesis in Bioinformatics and Systems Biology

JING GUO

Chalmers University of Technology
University of Gothenburg
Department of Computer Science and Engineering
Göteborg, Sweden, June 2014

Tissue-Specific Gene Expression Analysis
A large-scale meta-analysis of gene expression pattern using microarray data

JING GUO

Supervisors: DANIEL DALEVI
MARCUS BJARELAND

R&D Information
AstraZeneca
SE-431 83 Mölndal
Sweden
Telephone + 46 (0)31-776 1000

Examiner: OLLE NERMAN

Chalmers University of Technology
University of Gothenburg
Department of Mathematical Science
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

Department of Mathematical Science
Göteborg, Sweden June 2014

Abstract

An important problem in the early phases of drug-discovery and disease prognosis is to distinguish between genes that are ubiquitously expressed and genes that are preferentially expressed in one or a few tissues. Although several data sources and methods have been published explicitly for this purpose, it is still not evident how to retrieve these genes and how much confidence we can put in the results. We therefore investigate current data with the aim of answering queries, such as: *give me all genes that are preferentially expressed in e.g. liver*.

In this project we perform a meta-analysis on normal human tissues using a computational approach. Four large-scale microarray datasets across a broad range of normal human tissues were selected for the investigation of gene expression patterns. Methods from publications were implemented and tested on genes with known tissue-specificity. The parameters of the methods were optimized in an iterative procedure using these as training genes. Eventually we ended up using the *Decision Function* and *ROKU-SPM* – a method we modified and implemented. The results from separate datasets were combined using a rule-based score that represents the degree of specificity. Also, the *coverage* (indicating the presence of genes in the datasets) was used for assessing the confidence of the results. In total we produced a list of 27523 genes annotated with score and coverage. Among these, 1955 were predicted specific with the highest possible score.

The results were evaluated using three external databases: TiSGeD (microarray data), TiGER (EST data) and HPA (protein expression data). The 117 top candidates of specific genes, with highest score and highest coverage, were compared to these databases. 96.5% of our predictions were supported by at least one of the databases and agreed better with the consensus than the others.

Acknowledgement

This project is carried out within Research and Development Information at AstraZeneca Mölndal. The project was aimed to serve for an in-house integrated knowledge platform called PharmaConnect, which extracts, integrates and analyses knowledge to support systematic, evidence based decision making, regardless of discipline or content source.

I would like to give my sincere appreciation to those who supported and helped me through this master thesis project. My gratitude is beyond words, to my supervisor Daniel Dalevi and Marcus Bjäreland, for giving me this great opportunity to carry out a thesis work at AstraZeneca and for their endless support and encouragement throughout the completion of the project. I owe my deepest gratitude to Daniel Dalevi for his patience, motivation, enthusiasm and immense knowledge to drive me into the right direction, inspire me during the dark times and help me with his brilliant ideas. I would like to thank Marcus Bjäreland for providing breakthrough suggestions and the guide me throughout the project. I want to thank Olle Nerman, as my superviser and examiner, for his support and crucial suggestions.

I would like to thank Mårten Hammar and Lisa Öberg for their help with the vocabulary mapping and source data. Many thanks to the cortex team for providing valuable information and dynamic environment. I want to thank my friends, Saber Akhondi and Shanmukha Padmanabhuni, for their selfless help, as always.

Last but not least, I want to express my dearest gratitude and missing to my family. Without their support, encouragement and love, I would never become who I am, or do what I am capable to do.

Contents

1 Introduction

Target selection is the first step in pharmaceutical research and may open new paths to future development of drugs. Good quality of target selection will increase the success rate of drug discovery. A potential target can almost be anything that is related to the disease process [1]: small molecules, bioprocesses, pathways and network entities, etc. There are two common target selection procedures: the bioinformatics approach and the systems approach. The former aims at finding a druggable target among small molecules or proteins using computational methods, while latter uses clinical and *in vivo* information to analyze the pathways or networks and is less efficient on the large scale level and much more expensive. High-throughput biology in genomics and proteomics has dramatically created new possibilities for the bioinformatics approach. For example, large-scale analysis of human gene expression levels [2-4] can be used to improve target selection. It has been reported that genes with tissue-specific gene expression patterns are twice as likely to become drug targets compared to ubiquitously expressed genes [5]. Genome-wide expression profiling can be used to discover new disease genes and predict tissue specific expression patterns [6]. Among the more well-established approaches to study the transcriptome are: Microarrays, Expressed Sequence Tags (EST), Serial Analysis of Gene Expression (SAGE) and Massively Parallel Signature Sequencing (MPSS) technologies. Each of these has been used in the context of tissue-specificity.

Microarray technology has become one of the most popular since its rise in the early 1990s and is also the data used in this project. It allows the simultaneous measurement of the expression of thousands of genes [8-10] and there are a great number of datasets freely available on the Internet (e.g. GEO [11], BioGPS [12], ArrayExpress [13], etc). The methods are also well developed and tested which provides a solid background for further studies. There are also numerous studies related to tissue specificity in various species and method development. Here are some examples:

- In a study of 192 metabolic genes in fish, it was concluded that most of the differentially expressed genes among distinct tissues (76%) have a relationship to a distinct metabolic function [14].

- A compendium of normal human gene expression [15] identified 451 housekeeping genes and seven sets of tissue-selective genes among 19 different tissue types [16].

- A comprehensive analysis using both microarray and network topology detected 2374 housekeeping genes among 31 human tissues [5].

- AIC (Akaike's Information Criterion) is a non-threshold method using an outlier detection method to find tissue-specifically expressed genes [17].

- Sprent's non-parametric method, an outlier detection method, [18] was used to identify 2503 tissue-specific genes among 36 normal human tissues. They also performed a comparative analysis of genes in the cancer status, reporting that the profile of gene expression was significantly associated with their expression in cancer tissues.

- The ROKU method [19] uses both the Shannon entropy and a simplified AIC outlier detection method.

- There are also statistical methods such as the intersection-union test [20] and a Bayesian approach [2], where statistical models are applied to the raw data to quantify the degree of tissue-specificity.

EST data (Expressed Sequence Tags [21]) is another well established data type that allows high-throughput analysis of expression patterns. It has relatively low quality but there are several methods and data sources available for EST:

- TissueInfo [22], which is a knowledge-based method, reported an accuracy of 69% for identified tissue specificity against a set of 116 benchmark genes.

- ExQuest [23] first maps the EST data into a target sequence using MegaBLAST, and then uses the library annotation to compute the specificity of the corresponding gene.

- EST data are available in many public databases, such as Unigene, dbEST, etc.

SAGE (Serial Analysis of Gene Expression) [24] data and MPSS (Massively Parallel Signature Sequencing) [25] are less common compared to the previous two technologies. Both of them suffer from high cost and low efficiency, and are not suitable for large-scale analysis. Here are a couple of methods using SAGE and MPSS data:

- A method using SAGE data among 15 mouse tissues characterized constantly expressed genes with different expression levels and genes that are uniquely expressed in only one tissue [26].

- A methods using 400 pairs of MPSS allelic tags identified different regulation modes between tran- and cis-effects of expression [27].

- SAGE and MPSS data are available in Cancer Genome Anatomy Project (CGAP) [28] and Signature Sequencing data (the Ludwig Institute for Cancer Research) [25] respectively.

Some methods can be used for several data types:

- A method based on the Shannon entropy [29] can be applied to both microarray and EST data, which was used to discover the association between promoter features and tissue specificity.

- Another method based on the Dixon test and the gap distance of expression levels [4] can be applied on all kinds of transcriptomic data and even on protein expression data types.

As introduced above, the studies of gene expression pattern have obtained significant achievements in fields of tissue expression, differential rate of polymorphism and disease association etc. [30]. The study can be performed in different perspectives. For example, comparison of genes expressed in liver and lung; differences between normal thyroid and cancer thyroid tissues; the expression patterns of cell lines.

To clarify the scope of this project, first we present the definition of four types of gene expression patterns [2]:

1. A gene (e.g. CYP2C9) that is only expressed in one particular tissue [22] .
2. A gene (e.g. BDH1) that is expressed approximately the same level at all tissues except one.
3. A gene (e.g. CTRL) that is under or over expressed in a small group of tissues [17, 19].
4. A gene (e.g. CAPN10) that is ubiquitously expressed among all tissues [31].

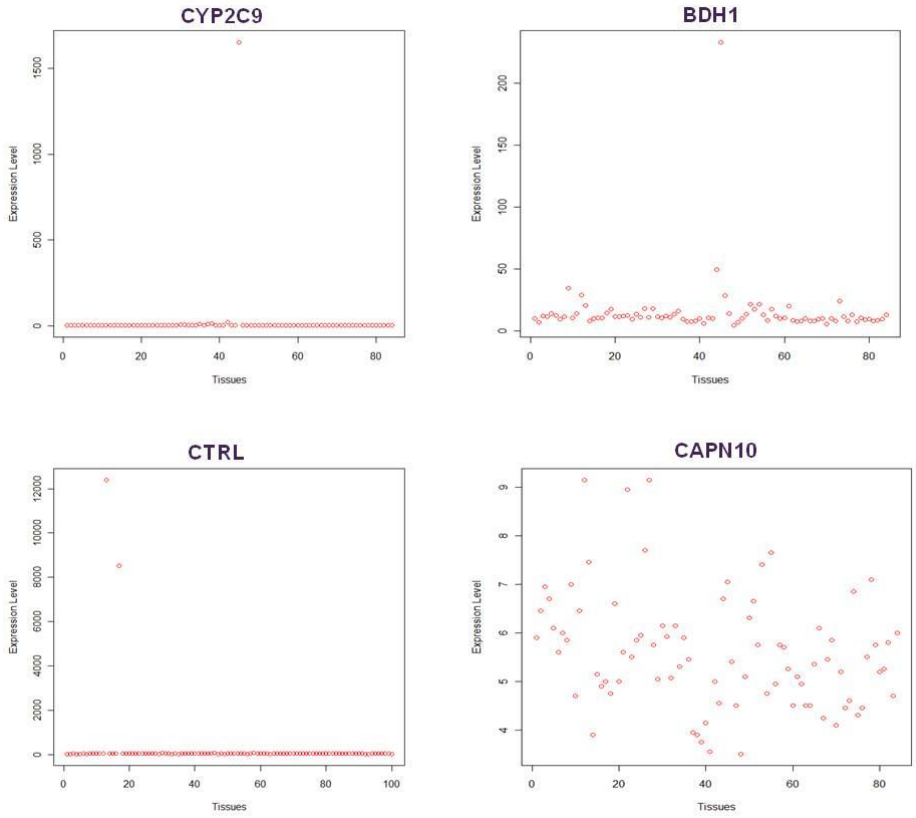

Figure 1 Gene expression patterns. Up-left: type 1; up-right: type 2; down-left: type 3; down-right: type 4. X-axis: tissue in arbitrary order; Y-axis: the relative expression level for each tissue.

The first and second types give the definition of *tissue-specific* genes. Type 1 is actually a special case of the second type. Type 3 is usually defined as *tissue-selective* [4], which in our project, only those with two selective tissues are discussed, i.e. *2-selective* genes. For clarification, the term *preferentially expressed genes* is used for both specific and 2-selective genes. Type 4 is also referred to as *Housekeeping genes* (HK genes) [5, 31] in some of the published papers but we differentiate between the HK and ubiquitously expressed genes (the former is a special case of the latter). Additionally, it is also possible to have two stereotypical expression situations: *up* and *down*, also refers as over- and under- expressed respectively [19]. The *up* type has higher expression intensity in a particular tissue compared to other tissues.

There is a similar concept of *differential expression* that is also used in the study of gene expression patterns. Although there are a few exceptions in the previously published paper, differential expression refers to the comparison between normal and disease tissues, for instance, the difference in expression between normal lung and cancer lung tissues. It is not within the scope of this project.

1.1 Objectives and organization

The project aims to predict specific and 2-selective genes in human tissues. The first task of this project is to provide appropriate methods to identify specific expressions among target tissues. The methods can be tested in typical examples of real data, and then applied on large scale data retrieved from the available data sources. As there are different datasets providing gene expression data, differences between data sources may lead to distinct results. Thus another challenge of this project is to evaluate the results derived from separate datasets, combine them and make predictions based on all sources.

This project is organized in the following procedure:

1. Data selection: select the data type and appropriate datasets.
2. Methods implementation: select or develop the detection methods on the data and implement them.
3. Optimization of methods: optimize the parameters in each method for each dataset using benchmark genes; guarantee the best performance of the methods.
4. Computation of the whole data in each datasets: applied the best methods on each datasets.
5. Combing the results for each dataset: meta-analysis and the integration of the result.
6. Evaluation of the result: compare the result with other tissue-specific databases.

1.2 Contribution

This project provides a meta-analysis on normal human tissues in a computational approach. Comparing to the previous studies of tissue-specific gene expression, which usually provide only one method on one dataset, this project optimized different methods (containing one derived method) based on the training of the parameters and the testing of the method performance, and then applied the selected methods on four large-scale datasets. The result from separate datasets are combined using a scoring algorithm, and a value of “coverage” indicating the presence of genes in this four datasets is also introduced for the confidence of the combined result. Finally, an integrated result containing 27523 genes is annotated with corresponding “score” and “coverage” indicating the level of specificity.

1.3 Structure of the report (outline of thesis)

This report consists of 5 parts:
1. Introduction (above): provides the background and motivation of this project;
2. Data: introduces the datasets that are used both for detection and analysis;
3. Methods: describes the data sources and methods that have been applied, along with the optimization of method-parameters and a scoring strategy for the integration of results;
4. Results: presents the results that are obtained by using the optimized methods on each data source, and introduces a method for the integration of results;
5. Conclusion and discussion: concludes a discussion of the possible aspects for the future work.

2 Data

This section gives a description of datasets and databases that have been used in this project. Generally, four large-scale microarray datasets are used for the analysis of gene expression pattern (Section 2.1). Three small sets of genes with known property of specificity are used for the training and testing of methods (Section 2.2.1). Three databases with preferentially expressed genes are used for evaluation of the results (Section 2.2.2).

2.1 Microarray data in use

The large scale genetics that has been tremendously developed in recent years enables the availability of millions of microarray data among all kinds of tissues, cell lines in normal or disease state. The selection of data becomes a crucial problem of this project. However, the idea of providing a general view of tissue-specific gene expression among normal human tissues narrows down the candidate data to a small group.

We choose four large scale datasets across a broad range of human tissues from three data sources (see Table 1). Two of them are the raw data with replicated tissue samples and absolute expression level and the other two are normalized expression level.

BioGPS is an online gene portal that can retrieve gene expression pattern and annotated information [12]. The data is downloadable through the Internet [32] and has a reference to GSE1133 in GEO. The datasets are, however, not identical (see below). The BioGPS data was used by GNF (Genomics Institute of the Novartis Research Foundation) [33] for the analysis of gene expression in human.

GEO (Gene Expression Omnibus) is a free database where a large amount of high throughput data (mainly microarray) are archived[11]. The database provides full annotations of the stored datasets and allows users to upload or download both the raw data and the annotations in an easy manner. The data are provided by different teams from all over the world and most of them are samples of particular sets of tissues, e.g. normal lung tissues and cancer lung tissues. There are few datasets that contain expression data from tissues of the whole human body. Among these, we selected the two datasets containing the largest number of probesets and the broadest range of tissues:

> **GDS3113**: “Various normal tissues” [1]. This dataset has 32878 probesets across 96 tissues with 3 replicate samples for each tissue. The data are not normalized and have a relatively high norisy level.
>
> **GDS596**: “Large-scale analysis of the human transcriptome (HG-U133A)” [Su AI, et al., 2004]. This dataset is also provided by GNF group (*The Genomics Institute of the Novartis Research Foundation*, which is the team supports the BioGPS database), with the same reference to GSE1133 in GEO. This indicates that the annotations of these two datasets are identical. However, the data from these two sources are completely different. The BioGPS data have been normalized by MAS5 and have 84 separate tissue samples. And GDS596 is raw data without any normalization and has 158 tissues samples with 2 replicate samples for each tissue. The data might come out from similar experiment, however, it is reasonable to study both since it is a good comparison for the preprocessed and non-preprocessed microarray data and a fair test of the methods.

GeAZr is a dataset licensed from GeneLogic by AstraZeneca. It consists of expression data from different donors but from the same tissue type (all determined as "normal" by pathologist) which were grouped and

averaged. It contains 44928 human genes expressed across 100 normal tissues, with complete data of type HG-U133 a and b.

Table 1 A summary of the selected datasets with some additional facts.

DATA SOURCES	NUMBER OF PROBESETS	NUMBER OF TISSUES	DATA TYPE	DATA	NOISY LEVEL
BioGPS	22283	84	HG-U133a	2004-03-19	Low
GDS596	22283	79	HG-U133a	2004-03-19	High
GeAZr	44928	100	HG-U133a, HG-U133b	2008-09-19	Low
GDS3113	32878	32	ABI Human Genome Survey Microarray	2007-06-04	High

As shown in Figure 2, data from BioGPS and GeAZr are less noisy than the data from GDS596 and GDS3113. However, as we want to bring in as many datasets as possible, we still keep these two datasets, and improve the result by modifying our methods.

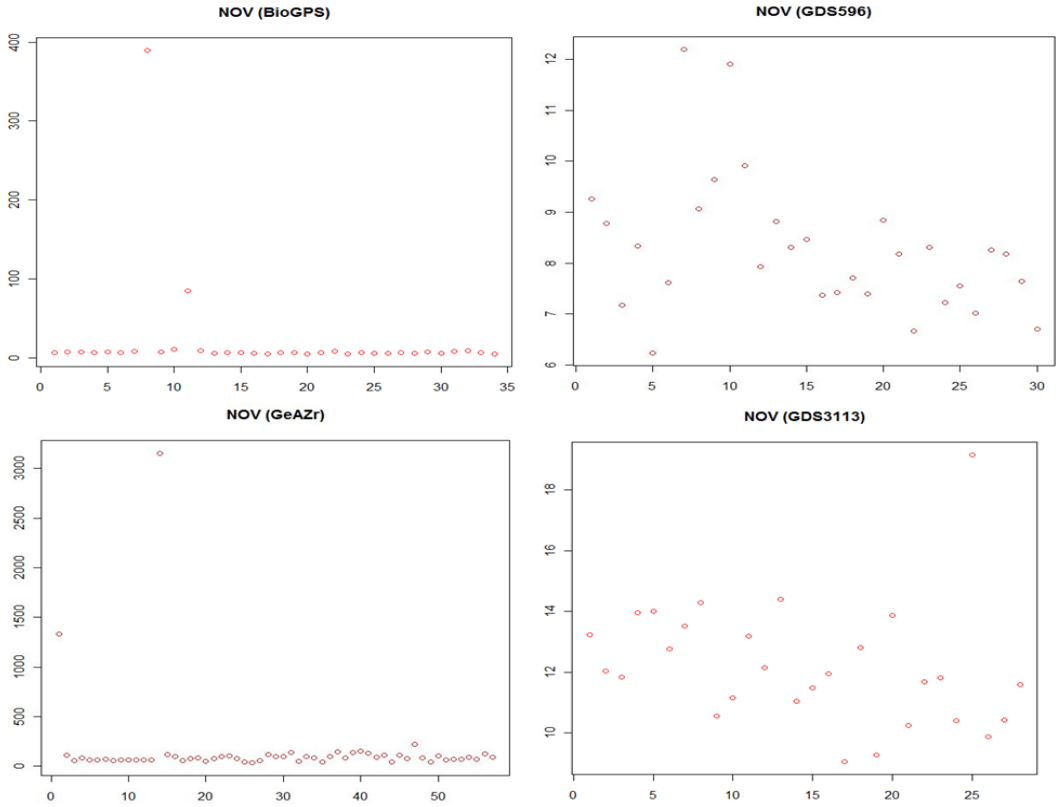

Figure 2 The expression distribution for gene NOV in four datasets.

2.2 Data for evaluation

To evaluate the methods used for the detection of specific genes, we choose three sets of genes with known specificity; and to evaluate the result of this project, three databases containing different types of expression information are used for comparison.

2.2.1 Standard genes

In order to optimize the parameters, 26 specific/selective genes are used for the training of the methods. These genes, referred to as "training set" (see in Table 8), are chosen from the supplemental information

of HugeIndex.org [16], under 'brain', 'kidney', 'liver', 'lung', 'muscle', 'prostate' and 'vulva' specific tabs.

Another two sets of standard genes, testing set 1 (see Table 9) and testing set 2 (see in Table 10), are chosen from the same sources to evaluate the result of each method. The data of testing set 1 is chosen from the liver-specific tab and testing set 2 is chosen from the 'housekeeping' group.

All the data of these standard genes are different and independent from the data for use in Section 2.1.

2.2.2 Databases for result comparison

There are a few databases available online and provide tissue-specific gene information. In this project, three databases representing three different kinds of data types are used for the comparison of our predicted result. The result of comparison is listed in Section 4.5.

TiSGeD (Tissue-specific Genes Databases) [34] uses microarray data and a statistic SPM to measure the specificity of genes. 123125 gene expression profiles across 107 human tissues, 67 mouse tissues and 30 rat tissues are measured in this database. Some literature records of tissue-specific genes are added in the database. An SPM cutoff is required to distinguish the specific genes from ubiquitously expressed genes. SPM is also discussed in this project (see Section 3.1.1).

TiGER (Tissue-specific Gene Expression and Regulation) [35] is a web database that gives a comprehensive information of human gene specificity using three types of data: the gene expression profile (EST), a combinational gene regulation (based on transcription factor binding sites) and cis-regulatory module (CRM). For the comparison of our predicted result, only the record from gene expression profile is used. It is also a good example for a comparison of EST data and Microarray data.

HPA (Human Protein Atlas) [36] uses high-resolution images to show protein expression profiles in 46 normal tissue, 20 cancer types, and 47 cell lines for human species. The gene-centric manner of HPA enables the comparison of proteomics data (antibody) and the genomic data (microarray). The expression intensity is marked as "level of antibody staining" with Strong, Moderate, Weak and Negative levels. Only genes marked with "Strong" are considered as specific/selective in this project. And only the 46 normal human tissues are used for comparison.

Table 2 Some facts about TiSGeD, TiGER and HPA.

DATABASES	DATA TYPE	SPECIES	NUMBER OF RECORDS	EXPRESSION PROFILE
TiSGeD	Microarray, Literature record	Human, Mouse, Rat	123125 probsets	107 human tissues, 67 mouse tissues, 30 rat tissues
TiGER	EST, TF, CRM	Human	19526 Unigene, 7341 TF pairs, 6232 CRMs for 2130 RefSeq genes	30 tissues
HPA	Antibody	Human	>700 antibodies	46 normal human tissue, 20 cancer types, 47 cell lines

3 Methods

This section introduces the main methods that are used in this project. Section 3.1 illustrates three established and one derived methods that are used for detecting specific genes. Section 3.2 describes the optimization of the methods in Section 3.1. Section 3.3 states a preliminary step before the meta-analysis, which is the vocabulary mapping among different data sources. And Section 3.4 explains the algorithm for integrating the result of all four datasets by calculating a combined score and coverage. Section 3.5 gives the details for the implementation of the methods in the previous sections.

3.1 The detection of preferentially expressed genes

In this section we derive a new method and describe known methods from the literature.

3.1.1 ROKU-SPM

ROKU and SPM are two methods that have been used to identify preferentially expressed genes. Both of them show advantage in certain situations, however, we define a new method ROKU-SPM by integrating these two methods which has a better performance in more general cases.

3.1.1.1 SPM

TiSGeD database [34] proposed a SPM method, which uses an *SPM* value (Specificity Measurement) to measure the specificity of each tissue in one gene. The *SPM* value is calculated as "the ratio of vector X_i's scalar projection in the direction of vector X_p against the length of X_p". As the projection can be calculated in many manners (absolute value, squared value, etc.), we use a squared projection in our method, which results in this formula:

$$SPM_{(g,t)} = \frac{x_t^2}{\sum_{t=1}^{N} x_t^2} ,$$

$SPM_{(g,t)}$ stands for the SPM value of gene g on tissue t, where N is the total number of tissues expressed in gene *x*, and x_t is the expression intensity of a gene in tissue *t*. The maximum value of $SPM_{(g,t)}$ is 1, which indicates the most specificity on tissue t. Contrarily, the minimum value of $SPM_{(g,t)}$ is 0, which indicates t is not expressed at all.

Here we present an example for the calculation of *SPM* for gene CYP2C9:

TISSUE	Liver	Appendix	Nerve	CNS	...	Total
SPM	0.99258	0.00016	0.00005	0.00005	...	1

In this case, the liver is detected as specific using the cutoff $SPM > 0.9$.

3.1.1.2 ROKU

Generally, the ROKU method [19] uses Shannon Entropy to measures the specificity of a gene expression, and an outlier detection method to identify the specific tissues if any exits.

Before calculating the Shannon Entropy, a one-step Tukey's biweight (T_{bw}), is used to improve robustness of the expression data:

$$x'_t = |x_t - T_{bw}|,$$

where x_t is the expression intensity of gene *x* in tissue *t*.

The Shannon entropy is calculated as

$$H(x) = -\sum_1^N p_t log_2 p_t,$$

where N is the total number of tissues, p_t is the relative expression of x_t for tissue t, defined as

$$p_t = \frac{x_t}{\sum_{t=1}^N x_t}.$$

If $H(x)$ is lower than a criteria, then gene is identified specific. A simplified AIC method [37] will be used to detect the outliers, which in our case are those specific tissues.

The AIC (Akaike's Information Criterion) [38] uses an equally treatment to all outliers, regardless of the number of outlier, the level of significance or the masking effects, which provides a good solution for the unbiased analysis of expressed tissues. To reduce the complexity of computation, Ueda [37] provide a simplified outlier detection method by using a statistic U_t:

$$U_t = \frac{1}{2} AIC = n log \hat{\sigma} - \sqrt{2} \cdot s \cdot \frac{log n!}{n},$$

where $\hat{\sigma}$ is the estimated standard deviation, s is the number of outlier candidates ($s = 0,1,2,3, \dots$), $n + s$ is the total number of observations (tissues in our case). For a set of observations, U_t is calculated for all combinations of (n, s), where the smallest U_t indicates the best option of outliers.

3.1.1.3 A Modified method: ROKU-SPM

Although there are good examples, the actual results of ROKU and SPM were not performing sufficiently on most of the training data compared to the other methods. In general, there are two problems:

1) For the ROKU method, there are cases where the entropies are incredibly low while a large number of outliers are detected.
2) When the data is noisy (GDS raw data), the difference between the entropy of specific and non-specific genes is hardly detectable. Similarly, for the SPM method, the *SPM* value of the specific tissue is not remarkable different to the other non-specific tissues.

For example, to illustrate the problems, we look at the probeset *214421_x_at* for gene CYP2C9. The figure below shows the expression distribution in BioGPS (left) and GDS596 (right). In BioGPS, although low entropy (0.527) and high SPM (0.99) supporting specificity for Liver, which is also easily caught by eye-browsing, the outlier detection method gives us 6 specific tissues (i.e. Problem 1). In GDS596, on the

other hand, we have high entropy (5.75) and low SPM (0.02) for Liver, this gene can hardly be identified as specific based on either the Entropy or SPM. The outlier detection method, however, correctly identifies Liver as a specific tissue (i.e. Problem 2).

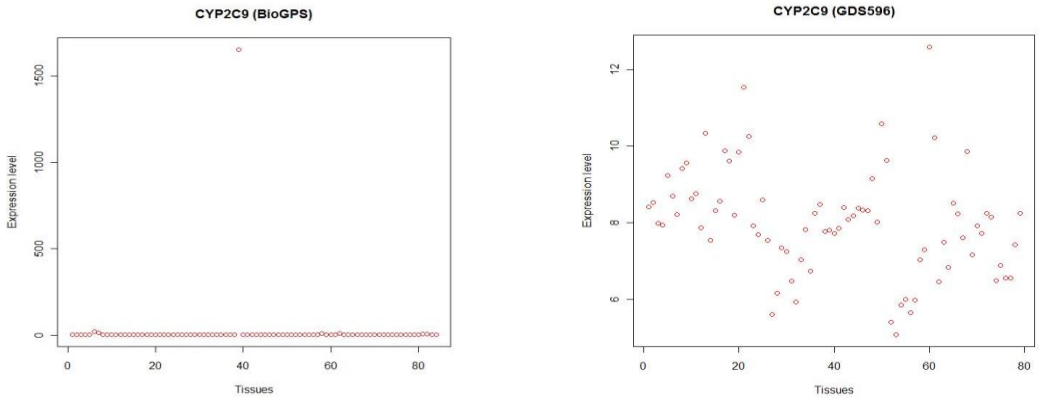

Figure 3 Expression distribution of gene CYP2C9 (*214421_x_at)* from BioGPS (left) and GDS596 (right).

One may argue that if the AIC based outlier detection method does not return appropriate number of specific tissues, why do not use another outlier detection method instead? This problem has been discussed by the provider of ROKU method, Koji Kadota, in another paper [39] by evaluating two outlier detection methods for microarray data. The AIC based method shows dominant advantage over the Sprent's non-parametric method [18], which to some extent answers this question.

To solve these two problems, we propose an improved method, which combines ROKU and SPM method, to resolve the two issues above. It is referred as ROKU-SPM method in this report. In this method, the *SPM* value is introduced as a parameter to the ROKU method, and it is used to classify the specific, 2-selective and ubiquitously expressed groups. The procedure of this method is shown in Figure 4.

A preferentially expressed gene must satisfy the following requirements:

1. The entropy is lower than E - the Entropy threshold.
2. The outlier with the largest value is greater SPM_1 – the first SPM threshold.

Similarly, the requirements for 2-selective genes:

1. The entropy is lower than E.
2. The outlier with the 2nd largest value is greater SPM_2 – the second SPM threshold

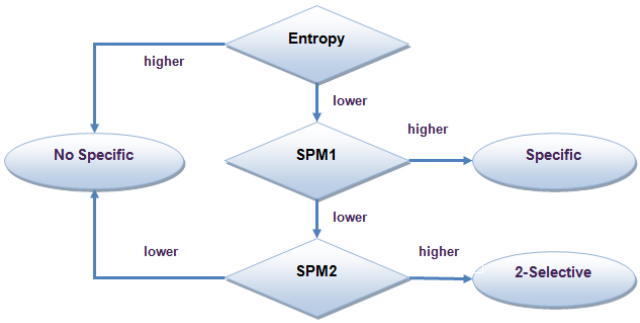

Figure 4 Flow chart of ROKU-SPM method

3.1.2 Decision function

This method gives a deterministic parameter (d) for gene specificity based on *gap* and a significance probability (*sp*). The gap indicates the absolute difference between the intensities of two tissues; the significance probability (*sp*) is calculated by a Dixon test:

$$sp = P[t \geq T_{critical}] = 1 - \int_0^{T_{critical}} F_{2,2n-2}(z)dz,$$

where $T_{critical}$ is the Dixon critical statistic, n is the total number of tissues, F is the standard statistical F distribution with $(2,2n-2)$ degrees of freedom.

One of the advantages of this method is that it takes the confidences of data sources into consideration, by using a adjusted value of sp:

$$\log(sp_{thresh}) = (n-2)\log(1-\lambda\tau),$$

where λ is the confidence level of sources (tissue sample in microarray data), and τ is a parameter based on sp.
However, tissue samples with replicates are considered to be equally trustable (GDS3113 and GDS596) and tissues without replicates does not really have this issue of confidence. So in this project, an uniform confidence level ($\lambda = 1$) are used in practice.

The indicator of gene specificity is calculated by a decision function:

$$d(g,s) = 1 - \left[(1-s)^{\alpha}(1-g)^{\beta}\left(\frac{\delta(1-g)+(1-\delta)(1-s)}{(1-g)+(1-s)}\right)^{\gamma}\right]^{\Phi},$$

where s and g are the variant of the *gap* and *sp*:

$$g = \begin{cases} 0, if\ gap \leq g_{thresh} \\ \frac{(gap-g_{thresh})}{(1-g_{thresh})}, if\ gap \geq g_{thresh} \end{cases},$$

and

$$s = \begin{cases} 0, & if\ log_{10}(sp_{adjust}) \geq log_{10}(sp)_{thresh} \\ 1, & if\ log_{10}(sp_{adjust}) \leq log_{10}(sp)_{\infty} \\ \frac{log_{10}(sp_{adjust}) - log_{10}(sp)_{thresh}}{log_{10}(sp)_{\infty} - log_{10}(sp_{thresh})}, & if\ log_{10}(sp)_{thresh} < log_{10}(sp_{adjust}) < log_{10}(sp)_{\infty} \end{cases}$$

$\alpha = \beta = \gamma = 1.5$ and $\delta = (\alpha+\beta+\gamma)^{-1} = 0.3$ are independent parameters chosen empirically by the authors of the original paper.

3.1.3 Bayesian approach

The Bayes factor [2] is used to measure the support for two hypotheses:

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

where μ_j denotes the expression level of tissue j in one gene, and J is the number of total tissues. And

$$H_2 : not\ H_1 \ .$$

H_1 states that the expression level of tissue 1 is larger than in the other tissues, which means that tissue 1 is specific for this gene.

In their modelling approach, a priori distributions under H1 and H2 are postulated from a joint model a priori distribution by conditioning on the hypothesis. This leads to a Bayes factor BF12 which is the ratio of the a priori and a posteriori odds of H1 (odds of H1 = ratio of the proabilities of hypothesis H1 and H2). They further advocate that one can, and should, use natural non-informative a priori distributions, so that the calculations are essentially based on the likelihoods. And $BF_{12} = 6$ implies that the support in the data for H_1 is 6 times as large as the support for H_2.

We use the value of BF_{12} as an index to measure the specificity of the gene on a certain tissue. And BF_{12} is the only parameter that has been optimized in the training section. It is notable that this method can be only applied to the datasets with replicates, which are GDS596 and GDS3113.

The application of this method is available in both R and matlab package[40], provided by the author of this method [2]. We consider it as a reliable source and use their implementation without further modification.

3.2 Optimization of parameters

The goal of the optimization process is to find the *best* parameters for each method on each datasets. An optimization function is introduced to quantify the performance and applied on the training gene set. The ideal result is that the method detected all training genes as one-specific or two-selective with correct tissues.

There are 3 parameters to be optimized in ROKU-SPM ($entropy$,$SPM1$, $SPM2$), 4 in Decision function (d,$s(min)$,$s(max)$, g) and 1 in Bayesian approach (BF_{12}).

3.2.1 Optimization function

A simple score (s) is used for evaluating the performance of a method (M_i) on the training data:

$$s(M_{i,\bar{p}}) = \sum_{g \in T} \left(\tau_{M_{i,\bar{p}}(g)} - 1\right)\left(\tau_{M_{i,\bar{p}}(g)} - 2\right),$$

where $\bar{p}$ is a parameter vector and τ is the number of tissues identified by method i given $\bar{p}$ and a gene g from the training set T. For each $\bar{p}$ of method i ($M_{i,\bar{p}}$), one score s is obtained. The function has a minimum *s* if the method identifies one or two tissues for each of the genes. Therefore, we regard the vector $\bar{p}$ that minimizes s as the best parameter set.

3.2.2 Training the parameters

The 26 training genes are used in this process. The similarity between the actual result and the expected result is measured by the optimization function in Section 2.5.

The procedure of training:

1) Constrain each parameter to an interval according to the distribution of the parameter itself. For example, the entropy of BioGPS data is between 0.045 and 6.110 (the first quantile, 25%, is 4.444). As we assume that the proportion of specific genes among all genes is no larger than 25%, we use the range from 0.045 to 4.444 as our preset scope to optimize. The same principle is applied to other parameters.

2) Run loops to estimate the combination of parameters. This step is repeated several times beginning with large steps on the whole interval to find approximate values. Then we use smaller steps to fine prune the parameters over specific intervals around those approximate values.

The parameters after training are listed in Table 7.

3.3 Vocabulary mapping

The four different data sources have their own tissue vocabularies each, namely BioGPS with 84 tissues, GDS596 with 79 tissues, GeAZr with 100 tissues, and GDS3113 with 31 tissues (see Table **1**). Although many of the tissues are shared in all datasets, it is necessary to reorganize the tissue terms to make them more comparable for the calculation of the combined score. In our approach we remove tissues that are out of interest (fetal tissues and cell lines) and group tissues that are functionally or literary similar. The list of tissues before and after grouping is shown in Table 3, and more detailed grouping information for each dataset is listed in the Appendix.

The principle of tissue grouping is to maximize the similarity of category in each dataset base on biological knowledge, so that the results from each method/dataset pair can be universally comparable. It is more focused on the literal similarity than biological sense. However, the taxonomy of tissues is arguable from distinct aspects. For example, should the 'Cerebellum' be classified under 'CNS'? The opinion may vary from different interests: for a general view of human tissue, the cerebellum is part of CNS; while for a particular scope of nerve system, probably the cerebellum should be individually analyzed. But based on our principle above, since the "cerebellum" exists in 3 datasets (i.e., BioGPS, GDS596 and GeAZr) out of 4, it is more reasonable to keep it separated from CNS. The same issues can be also applied to other tissues, and may become a potential interest in the future work. A further discussion of this issue will be enclosed in the Discussion Section.

Table 3:a Vocabulary Mapping. The grouping of functionally and literary similar tissues. The original term is mapped onto the new term. (part 1)

New term	Original term	New term	Original term
CNS	Brain	**Nerve**	Nerve
	Frontal cortex		Ciliary ganglion
	Occipital cortex		Dorsal root ganglion
	Parietal lobe		Superior cervical ganglion
	Amygdala		Trigeminal ganglion
	Temporal lobe	**Pineal**	Pineal night
	Prefrontal cortex		Pineal day
	Temporal cortex	**Uterus**	Uterus
	Parietal cortex		UterusCorpus
	Corpus callosum	**Testis**	Testis
	Locus ceruleus		TestisSeminiferousTubule
	Caudate nucleus		TestisGermCell
	Putamen		TestisIntersitial
	Nucleus Accumbens		TestisLeydigCell
	Globus pallidus	**Salivary gland**	Parotid gland
	Subthalamic nucleus		Salivary gland
	Substantia nigra	**Vas deferens**	Vas deferens
	Medulla oblongata	**Articular surface of bone**	Articular surface of bone
	Nucleus basalis of Meynert	**Bone structure**	Bone structure
	Amygdaloid nucleus	**Meniscus of joint**	Meniscus of joint
	Hippocampus	**Tendon**	Tendon and tendon sheath
	Hypothalamus	**Soft tissue**	Soft tissues
	Pulvinar	**Omentum**	Omentum
	Thalamus	**Skeletal muscle**	Skeletal muscle
	Cingulate gyrus	**Smooth muscle**	Smooth muscle
	Cingulate cortex	**Spinal cord**	Spinal cord
	Caudatenucleus	**Placenta**	Placenta
	Olfactory bulb	**Bronchus**	Bronchus
	Pons	**Larynx**	Larynx
	Entorhinal cortex	**Lung**	Lung
	Pituitary gland	**Trachea**	Trachea
	Red nucleus	**Bladder**	Bladder
	White matter of occipital lobe	**Kidney**	Kidney

Table 4:b Vocabulary Mapping. (part 2)

New term	Original term	New term	Original term
Cerebellum	Cerebellum	**Ureter**	Ureter
	CerebellumPeduncles	**Urethra**	Urethra
Vessel	Aorta	**Mammary gland**	Mammary gland
	Artery	**UHR**	Universal Human Reference
	Abdominal aorta	**Skin**	Skin
	Ascending aorta	**Epididymis**	Epididymis
	Blood vessel	**Cardiac muscle**	Cardiac muscle
	Coronary artery	**Prostate**	Prostate
	Vein	**Seminal vesicle**	Seminal vesicle
Heart	Heart	**Vagina**	Vagina
	Left atrium	**Vulva**	Vulva
	Left ventricle	**Lymph node**	Lymph node
	Right atrium	**Spleen**	Spleen
	Right ventricle	**Thymus**	Thymus
Bile	Common bile duct	**Tonsil**	Tonsil
	Gallbladder	**Blood**	White blood cell
Liver	Hepatic duct	**Brest**	Brest
	Liver	**Esophagus**	Esophagus
Pancreas	Pancreas	**Rectum**	Rectum
	Pancreas Islet	**Stomach**	Stomach
Small intestine	Duodenum	**Tongue**	Tongue
	Ileum	**Thyroid**	Thyroid gland
	Jejunum	**Cervix**	Cervix
	Small intestine	**Endometrium**	Endometrium
Adrenal	Adrenal cortex	**Fallopian tube**	Fallopian tube
	Adrenal gland	**Myometrium**	Myometrium
Adipocyte	Perirenal fat	**Ovary**	Ovary
	Adipose tissue	**Appendix**	Appendix
	Adipose tissue of breast	**Colon**	Colon

3.4 Scoring strategy

The four datasets used in this project are selected from different microarray databases with distinct platforms of probeset annotation. As the original data is recorded under probeset names, the first challenge of the result integration is the mapping from probesets to a uniform set of gene IDs (Section 3.4.1). Then a general measurement of specificity for each gene is used across different datasets (Section 3.4.2). Two confidence indices are introduced to classify the integrated result, namely the coverage of sources (Section 3.4.2) and the score of specificity (section 3.4.3).

3.4.1 Map probesets to genes.

The general information of probesets from each dataset is listed in Table 4.

Table 4 Mapping from probesets to genes.

DATA SOURCES	NUMBER OF PROBESETS	NUMBER OF MAPPED GENES	NUMBER OF UNMAPPED PROBESETS	ANNOATTION SOURCE
BioGPS	22283	13897	871	HGU133a.db (Bioconductor)
GDS596	22283	13897	871	HGU133a.db (Bioconductor)
GeAZr	44928	22771	4473	HGU133plus2.db (Bioconductor)
GDS3113	32878	16649	16752 (Removed)	GPL2986 (GEO)

Data from BioGPS, GDS596 and GeAZr share the same annotation from Affymetrix HGU133, which is available in R packages (Bioconductor [41]). There is a small proportion in each of these 3 datasets that lack of annotated information. However, since the Affymetrix HGU133 data is one of the most frequently used microarray data and the unmapped probesets might also be comparable between these three datasets, we kept the records that are unable to be annotated by gene IDs and left them as record of genes (Figure 5).

Data from GDS3113 use a particular annotation system from ABI Human Genome Survey Microarray, which is accessible in GEO platform GPL2986. There are 16752 out of 32878 probesets, which is more than half, do not have annotated information. They are not as valuable and comparable as HGU133 data since the GDS 3113 is the only dataset using this platform, and it will be problematic if we keep so much data without annotation, so we removed these 16752 for the integration analysis, and leave them to the future study.

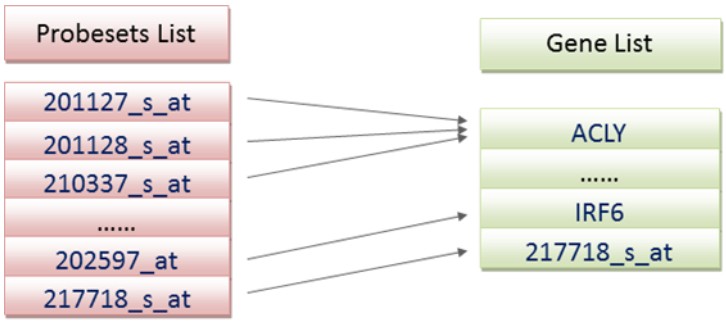

Figure 5 A demonstration of the mapping from probesets to gene symbol

3.4.2 The combination of method/dataset pairs

After the mapping of the probesets, data from different sources are comparable by using gene symbols. The overlap of gene among four datasets is shown in Figure 6. BioGPS and GDS596 share exactly the same gene set, so they are represented by one ellipse (in purple).

BioGPS and GDS596 share completely the same gene set (HGU133a, see Section 2.1); while GeAZr and GDS3113 both have large unique gene sets of their own.

As an important confidence index for the cross dataset analysis, the concept "coverage" of a gene is defined as the amount of datasets that contain this gene. It is generally believed that with the same score, genes that have higher coverage is more convinced to be specific than those with lower coverage. The number of genes in each coverage is listed in Table 5.

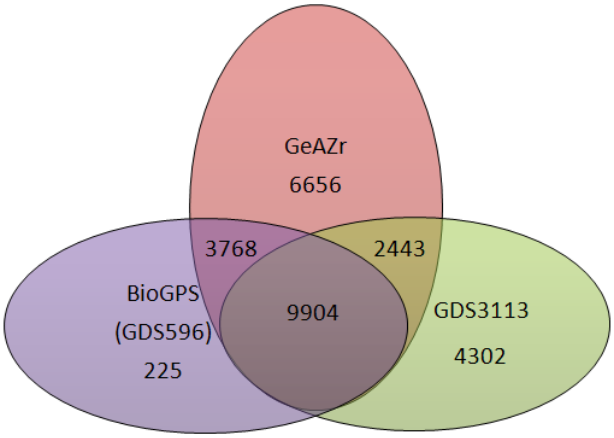

Figure 6 The overlap of genes among datasets.

Table 5 Number of genes for each overage.

COVERAGE	NUBMER
4 out of 4	9904
3 out of 4	3768
2 out of 4	2443
1 out of 4	11408
Total	**27523**

3.4.3 Scoring algorithm

The scoring algorithm consists of 2 parts: an *inner score* $i_s(T)$ inside one dataset; and a *total score* $t_s(T)$ combining all datasets, where T refers to tissues that are detected specific.

It is not rare that results of probesets from the same gene conflict to each other, and there is no universal rule to judge which is more accurate. In this case, we use a *majority vote* to capture as much information as we can: if more than half of the probesets indicates the same tissue as specific/2-selective, we regard the ubiquitously expressed probesets as *non-informative*. For example, if three out of five probesets are detected as liver specific, and the other two are not specific to any tissues, then we believe that there is enough evidence that this gene liver-specific, and give a full support with $i_s(T) = 1$ to liver specificity.

The score is calculated in the following steps (step 1 and 2 for inner score and step 3 for total score):

1. If at least 50% of the probesets are either specific or 2-selective and contains the same tissue T, remove all probesets indicating ubiquitously expressed – we regard them as *non-informative*.

2. If at least 50% of the remaining probesets are either specific for the same tissue (T) or 2-selective for the same two tissues (T_1, T_2), then $i_s(T) = 1$, for that tissue, or $i_s(T_1) = i_s(T_2) = 0.5$ for the two tissues. Otherwise, $i_s(T)$, for each detected tissue, will be the average of the frequency of the tissue over the probesets.

3. The *total-score*, $t_s(T)$, is the average over the inner-scores, hence constrained between 0 and 1, over all datasets.

Here we present two examples to explain the score algorithm:

Example 1: VGLL1

There are three probesets representing for the gene VGLL1 in BioGPS, GDS596 and GeAZr data, and one probeset in GDS3113. First, the inner score $i_s(T)$ is calculated for each data source separately. Both in BioGPS and GeAZr, two probesets out of three detect the gene as specific in Placenta, so the probesets 215730_at (marked with red circle) are regarded as non-informative (rule 1, original paper) and at least 50% of the remaining probesets indicate Placenta so the inner score will be one (rule 2, original paper). In GDS596, there are no non-informative probesets and no majority of tissues, so the score is obtained by averaging the result of each probeset (rule 2, original paper). In GDS3113, the inner-score is one (rule 2, original paper). Second, the total score $t_s(T)$ of a tissue is the average of the all inner scores.

GENE: VGLL1	PROBESETS	SPECIFIC TISSUES	INNER SCORE	COMBINED SCORE
BioGPS	215729_s_at	Placenta	$i_s('Placenta') = 1$	$t_s('Placenta') = \frac{(1 + 1/3 + 1 + 1)}{4} = 0.83$
	215730_at	-		
	205487_s_at	Placenta		
GDS596	215729_s_at	Placenta	$i_s('Placenta') = 1/3$	$t_s('Skeletal\ muscle') = \frac{(1/6)}{4} = 0.042$
	215730_at	Skeletal muscle, Nerve	$i_s('Skeletal\ muscle') = 1/6$	$t_s('Nerve') = \frac{(1/6)}{4} = 0.042$
	205487_s_at	-	$i_s('Nerve') = 1/6$	
GeAZr	215729_s_at	Placenta	$i_s('Placenta') = 1$	
	215730_at	-		
	205487_s_at	Placenta		
GDS3113	147124	Placenta	$i_s('Placenta') = 1$	

In conclusion: VGLL1 is predicted as Placenta-specific with *high* support, i.e. $0.8 \leq t_s(T) < 1$.

Example 2: RAPGEF5

The gene RAPGEF5 in BioGPS, GDS596 and GeAZr data has 2 probesets respectively, but none exists in GDS3113. The inner score in each data source is: BioGPS, one probesets out of two indicated that the gene is specific in spinal cord, so the probeset 204680_s_at (marked with red circle) is regarded as non-informative (rule 1, original paper). Similarly in GDS596 and GeAZr, RAPGEF is detected as 2-selective and the non-informative probesets are removed (rule 1, original paper) and the inner score obtained by the average (rule 2, original paper). Then, the total score is obtained by averaging the tissues scores among the 3 datasets.

GENE: RAPGEF5	PROBESETS	SPECIFIC TISSUES	INNER SCORE	COMBINED SCORE
BioGPS	204680_s_at	-	i_s('Spinal cord') = 1	t_s('Spinal cord') = 0.67 t_s('CNS') = 0.33
	204681_s_at	Spinal cord		
GDS596	204680_s_at	-	i_s('Spinal cord') = 0.5	
	204681_s_at	Spinal cord, CNS	i_s('CNS') = 0.5	
GeAZr	204680_s_at	-	i_s('Spinal cord') = 0.5	
	204681_s_at	Spinal cord, CNS	i_s('CNS') = 0.5	
GDS3113	/	/	/	

In conclusion: RAPGEF5 is predicted as 2-selective for the tissues Spinal cord and CNS with *medium-strong* support, i.e. $(t_s(T_1), t_s(T_2)) \geq (0.3, 0.3)$.

3.5 Implementation of methods

This section provides a browse of application of methods above, a more detailed instruction can be found in Appendix.

3.5.1 R

R is a free software environment for statistical computation and graphics[41]. Except for the various statistical functions in the core program, R has plentiful extension packages for disparate demands, e.g. the BayesianIUT package [40] for Bayesian method. Moreover, R has a user-friendly interface and simple syntax which enables user to define their own functions easily. The implementation of ROKU-SPM, Decision function and Optimization procedure is completed in R with self-defined functions (see Table 6).

Table 6 The Implementation of methods in R and Perl. (ROKU-SPM, Decision, BF and Optimization are programmed in R; the vocabulary mapping and scoring algorithm are programmed in Perl).

FUNCTION	PARAMETERS	OUTPUT	SUB FUNCTIONS	RUNING TIME
ROKU-SPM	entropy, spm1, spm2	number of specific tissues	outlier, tukey-biweight, shannon.entropy, spm,	medium
Dtest	d, g, s(min), s(max)	number of specific tissues	DECISION	short
BFsummary (Bayes Factor)	BF	BF value	BF	medium
Tmap	number of specific tissues, dataset number	formatted output with names of specific tissues	Names	short
Optimization	scope, interval	score matrix	Opscore, Oploop	long, repeatedly
Vocabulary mapping	four vocabularies	one combined vocabulary	read, map	short
Scoring algorithm	result from four datsets	combined result	inner score	medium
			total score	short

Functions in R can be saved in separate files and invoked by the command "source". Each sub function is an independent module and can be recalled in different functions. The detail of the sub functions are listed in the Appendix.

ROKU-SPM uses the thresholds for $Entropy$, SPM_1, and SPM_2 as input parameters, and outputs the number of specific tissues with value 0, 1 and 2 indicating "ubiquitously expressed", "specific" and "2-selective" respectively. The sub functions are used as module and sourced in the function "ROKU-SPM.R". A formatted output can be created by the function "Tmap" with listed names, corresponding gene names number of specific tissues, names of specific tissues for each probeset (see figure below):

```
> bg_all_list[1:5,]
      PROBE SYMBOL SP_NR    TISSUE.1 TISSUE.2
1 1007_s_at   DDR1     0        <NA>     <NA>
2   1053_at   RFC2     0        <NA>     <NA>
3    117_at  HSPA6     1 BTO:0000089     <NA>
4    121_at   PAX8     1 BTO:0001379     <NA>
5 1255_g_at GUCA1A     1 BTO:0001067     <NA>
```

Figure 7 The formatted output from R for BioGPS data using ROKU-SPM methods. The first 5 probesets are displayed in the figure: SP_NR is the number of specific tissues with maxim value of 2. TISSUE.1 and TISSUE.2 are the specific tissue names (using internal id in AstraZeneca).

Similarly, the decision function (Dtest.R) has four parameters as input value and the number of specific tissues as output value. "DECISION.R" gives the output value for one gene and "Dtest.R" summaries the result for a whole input datasets. "Tmap" is also used for formatting the output.

The output of the Bayesian method (BFsummary) is a little different, as it reports the BF value for each tissue in each gene, and only tissues with BF value above the input threshold are considered as specific.

The optimization function is scored by function "Opscore.R", and looped by "Oploop.R". The method should start with a broad scope and large interval, and narrow down gradually.

Additionally, two R packages, hgu133a.db and hgu133plus2.db, are used also in the mapping from probesets to gene symbols (see Section 3.4.1).

3.5.2 Perl

Perl is originally designed for text manipulation, and then broadened as a general purpose programming language. It is remarkably efficient in practice. The vocabulary mapping (Section 3.3), which is actually a text mining problem, and the scoring method for the combination of four method/dataset pairs (Section 3.4), which requires efficiency and text extraction, are implemented in Perl.

The inner score of the Scoring algorithm first remove the non-informative probesets (see Section 3.4.3), then average the result of probesets for the same gene. The total score maps the genes from different datasets, marks them with “coverage”, and integrate them with an average score.

Perl is also involved in the statistic for the analysis of results. For example, the statistic of specific tissue/tissue pairs among the 1955 most specific genes (Section 4.6).

4 Results

4.1 Training and optimization

We used the training set of 26 *known* specific/selective genes (see Section 2.2.1) to optimize the parameters of all method/dataset combinations. As the Bayesian Approach can be only applied to data with replicate samples (GD3113 and GDS596), there are 10 method/dataset pairs that have been optimized. As we mainly are interested in genes that are specific to one tissue or selective to no more than two tissues, an optimization function in Section 3.2.1is used to quantify the performance of parameter. The result of optimum parameters for each method on four datasets is listed in Table 7. In Table 8 we listed the identified tissues for the best parameters for all combinations method/dataset pairs. Grey color indicates the correct output. We saw that most methods performed well and in a few cases all methods could identify the correct tissue given by HugeIndex.org (tissue colored by grey). For example, the gene SFTPC was correctly identified as liver specific by all methods. Although there are cases like PMP22 that no method detect the correct tissue, in most cases, majority of the methods tend to report the same tissue.

Table 7 The parameters chosen after optimization.

	ROKU-SPM			DECISION FUNCTION				BF
	Entropy	SPM1	SPM2	d	s (min)	s (max)	g	
BioGPS	3.5	0.65	0.4	0.5	-4	-12	0.5	500
GDS596	4.45	0.055	0.03	0.5	-1	-3	0.5	500
GeAZr	4.93	0.41	0.25	0.5	-5	-14	0.5	500
GDS3113	4.35	0.066	0.06	0.5	-1	-3	0.5	500

Note: For the ROKU-SPM method, the thresholds for *entropy*, *SPM1* and *SPM2* are optimized; for the decision function method, the *s*, *g* and thresholds for *d* are optimized; for Bayes factor method, only the threshold for *BF* is optimized.

Table 8 Resulting tissues are shown when applying the chosen methods with the optimized set of parameters to the training data. '-' indicates that no specific tissue is identified, '/' indicates that the gene is not in the dataset. The methods applied are ROKU-SPM (RS), Decision Function (DEC), and Bayes Factor (BF). GREY color shows that the detected result is exactly same as the HugeIndex database, and the RED color shows that the detected result is partially same as the HugeIndex database.

Data	BioGPS		GDS596			GeAZr		GDS3113			Tissue Names
Method	RS	DEC	RS	DEC	BF	RS	DEC	RS	DEC	BF	
Gene	Detected tissues										
FXYD2	T1	T1	T1	T1	T1	T1,T2	T1	T1,T2	-	T1	T1=Kidney,T2=Bile
PAX8	T	T	-	-	-	T	T	T	T	T	T=Thyroid
GPX3	-	-	-	T	T	-	-	-	-	T	T=Kidney
AQP2	T1	T1	T1	T1	-	T2	T2	T1	T1	T1	T1=Kidney, T2=Vasdeferens
HABP2	T1	T1	T1	T1	T1	T1	T1	T1,T2	T1	-	T1=Liver, T2=Bile
SAA4	T1	T1	T1	T1	T1	T1	T1	T1,T2	T1	T1	T1=Liver, T2=UHR
CPN2	T1	T1	-	T1,T2	-	T1	T1	T1	T1	T1	T=Liver,T2=Nerve
ASGR1	T1	T1	T1	T1	T1	T1	T1	T1,T2	T1	T1	T1=Liver, T2=UHR
LIPC	T	T	T	T	-	T	T	T	T	T	T=Liver
SFTPC	T	T	T	T	T	T	T	T	T	T	T=Lung
SFTPB	T	T	T	T	-	T	T	T	T	T	T=Lung
RNF5	T1	T1	-	-	-	-	T2	/	/	/	T1=Heart,T2=Bile
CLDN5	T	T	-	T	T	-	-	-	-	T	T=Lung
PMP22	T1	-	-	-	-	-	-	-	-	T2	T1=CNS, T2=Spinal cord
MYOC	T1	-	-	-	-	T2,T3	-	-	-	T1	T1=Retina, T2=Tendon, T3=Meniscus of joint
TPM1	T1	T1	-	T1,T2	-	T1,T2	-	T1,T2	T1,T2	-	T1=Heart,T2=Skeletal muscle
MYH8	-	-	T4	-	-	T2,T3	-	-	T1	T1	T1=Kidney, T2=Skeletal muscle, T3=Bone structure, T4=Nerve
SGCA	-	T3	T1,T3	T1,T3	-	T1	T1	T1,T2	T1,T2	-	T1=Skeletal muscle, T2=Pancreas,T3=Heart
KLK3	T1	T1	T1	T1	T1	T1	T1	T1,T2	T1,T2	T1	T1=Prostate,T2=Salivary gland
ACPP	T	T	T	T	T	T	T	T	T	T	T=Prostate
MSMB	T1	T1	T1,T2	T1,T2	-	T1,T2	T2	T1,T2	T1,T2	T1	T1=Prostate, T2=Trachea
FLNB	-	-	-	-	-	-	-	-	-	T	T=Prostate
KLK2	T	T	T	T	-	T	T	T	T	T	T=Prostate
KRT10	T1	T1	T1	T1	T1	T1,T2	T1	/	/	/	T1=Skin, T2=Vulva
S100A7	T2	T2	T1,T2	T1,T2	T2	T1,T3	-	T1,T3	T1	T1	T1=Tonsil, T2= Tongue,T3=Larynx
LOR	T	T	T	T	T	T	T	T	T	T	T=Skin

4.2 A clustering analysis

To evaluate the similarity of the results, we clustered the result from each method/dataset pairs using a distance measurement. Additionally, we also added the databases (TiSGeD, TiGER and HPA) as external sources to be compared with. A standard hierarchical clustering was used with distance value ($d_k(i,j)$) defined in a simple manner. For each gene, if the result is the same, the distance between them will be 0; if partially same (share at least one same tissue), then 0; and if not the same at all, the distance will be 1:

$$d_k(i,j) = \begin{cases} 0, \text{ if the result of method } i \text{ and } j \text{ are the same} \\ 0.5, \text{ if the result of method } i \text{ and } j \text{ are partially the same} \\ 1, \text{ if the result of method } i \text{ and } j \text{ are different} \end{cases}$$

Where *k* is the number of gene, and *i*, *j* stands for the methods. The total distance between two methods is the sum of distances for the training genes ($N = 26$).

$$D(i,j) = \sum_{k=1}^{N} d_k(i,j) \ .$$

The distance matrix is formed by $D(i,j)$ and the R function *hclust* is used to perform the clustering (Figure 8).

As shown in Figure 8, the result from the same dataset rather than the same method, trend to be clustered together. This to some extent infers that the data matters more than the methods, and one representative for each dataset would be sufficient to show the result. Thus a selection procedure is applied on the method/dataset pairs in the next Section and eventually one *best* method for each dataset is chosen to be used in the whole datasets.

Additionally, the TiGER and HPA appears to be distant from the other ones. It is predictable because of the missing genes in their databases, and the different types of data they based on (EST and protein antibody).

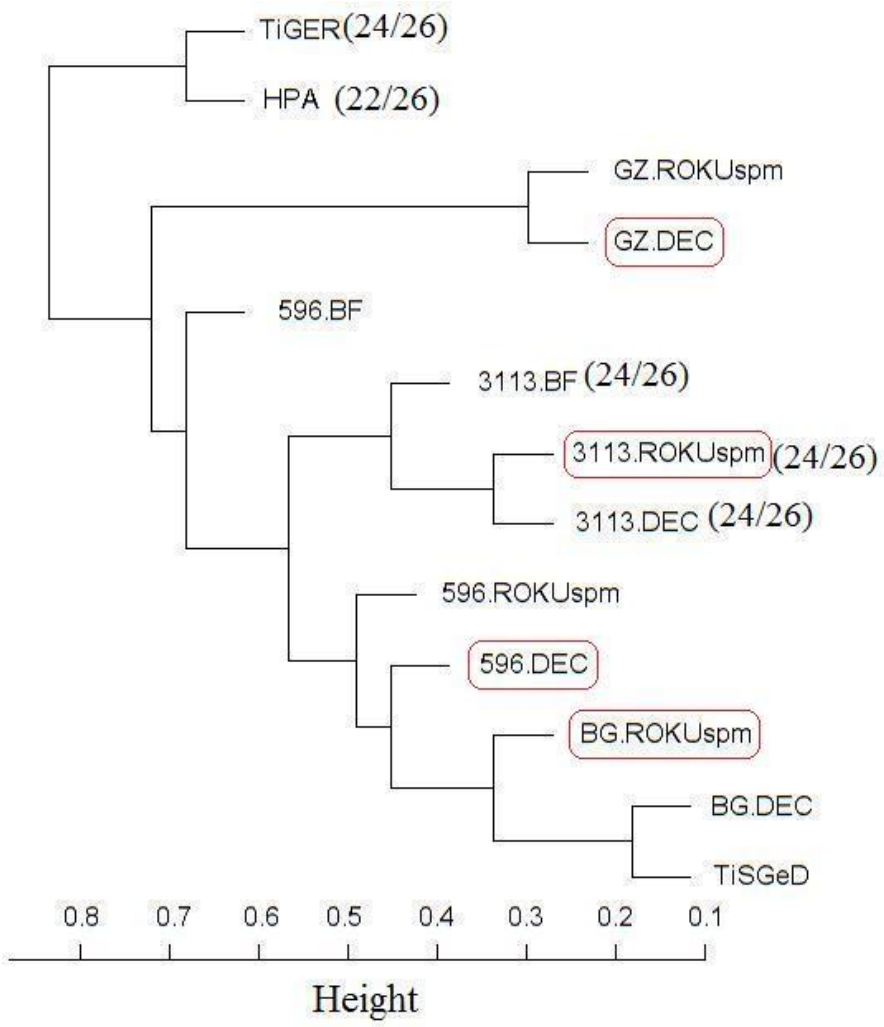

Figure 8 Clustering on 26 training genes among different method/dataset pairs and the three databases (HPA, TiSGeD and TiGER). 'BG' is BioGPS dataset, '596' is GDS596 dataset, 'GZ' is GeAZr dataset, '3113' is GDS3113 dataset, 'DEC' is Decision Function method, 'BF' is Bayes Factor method. The red rounded rectangles indicate the selected method/dataset pairs based on their performance on the testing gene sets.

4.3 The selection of methods for each data sources

The parameters obtained from the training are also applied to the two testing gene sets (data in Section 2.2.1). Table 9 shows the output from testing set 1 of 10 liver-specific genes and Table 10 shows the output from testing set 2 of 10 housekeeping genes (ubiquitously expressed genes).

An error evaluation is employed on the results of two testing sets in Table 11, with four parameters: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). TP and FN are derived from testing set 1, and TN and FP are derived from testing set 2, where TP+FN=10 and TN+FP=10. The best possible performance is TP=TN=10 where all *true* are found and all *false* are rejected. We selected the method that has best performance on each datasets based on these numbers,

which is ROKU-SPM for BioGPS (TP=6, TN=10, FP=0, FN=4) and GDS3113 (TP=10, TN=10, FP=0, FN=0) and decision function for GeAZr (TP=9, TN=10, FP=0, FN=1) and GDS596 (TP=3, TN=10, FP=0, FN=7). Since there is not much difference, it might be nature to ask why not use the same methods on all datasets. Well, on the small dataset of 10 genes, the slight difference might not causing problems, but on the whole data with tens of thousands genes, the difference will be significant. Besides, the use of different method may enrich the analysis and avoid the bias by applying only one method on all data.

Table 9 Resulting tissues are shown when applying the chosen methods with the optimized set of parameters to the testing data 1 (liver-specific genes). '-' indicates that no specific tissue was identified. The methods applied are ROKU SPM (RS) and Decision Function (DEC). GREY color shows that the detected result is exactly same as the HugeIndex database, and the RED color shows that the detected result is partially same as the HugeIndex database.

DATA	BioGPS		GDS596			GeAZr		GDS3113			Tissue Names
Method	RS	DEC	RS	DEC	BF	RS	DEC	RS	DEC	BF	
Gene	Detected tissues										
F12	T	T	T	T	T	T	T	T	T	T	T=Liver
CYP2C18	T1	-	-	-		T1,T2	-	T1,T2	-	T1	T1=Liver,T2=Small intestine
CYP2C9	-	-	T3	T3	T3	-	T1	T1,T2	T1	T1	T1=Liver,T2=Small intestine,T3=Nerve
ITIH2	T1	T1	-	T1,T2	T1,T2	T	T	T1,T3	T1,T3	T1,T3	T1=Liver, T2=Nerve, T3 = UHR
C8G	T	T	T	T	T	T	T	T	T	T	T=Liver
CYP3A7	-	-	T3	-	-	T	T	T1,T2	T1,T2	T1	T1=Liver,T2=Small intestine, T3=Nerve
TDO2	T	T	-	-	-	T	T	T	T	T	T=Liver
CRP	T2	T2	-	-	-	T1	T1	T1	T1	T1	T1=Liver, T2=Pancreas
MBL2	T	T	-	-	-	T	T	T	T	-	T=Liver
MTP	T2	T2	-	-	-	T2	T1,T2	T1,T2	T1,T2	-	T1=Liver,T2=Small intestine

Table 10 Resulting tissues are shown when applying the chosen methods with the optimized set of parameters to the testing data 1 (liver-specific genes). '-'indicates that no specific tissue was identified. The methods applied are ROKU SPM (RS), and Decision Function (DEC). GREY color shows that the detected result is exactly same as the HugeIndex database.

DATA	BioGPS		GDS596			GeAZr		GDS3113			Tissues Names
Method	RS	DEC	RS	DEC	BF	RS	DEC	RS	DEC	BF	
Gene	Detected tissues										
RPL29	-	-	-	-	-	-	-	-	-	-	
H3F3B	-	-	-	-	-	-	-	-	-	T	T=Testis
RPS26	-	-	-	-	-	-	-	-	-	-	
BAT1	-	-	-	-	-	-	-	-	-	-	
SURF1	-	-	-	-	-	-	-	-	-	T	T=Salivary gland
RPL8	-	-	-	-	T	-	-	-	-	-	T=Heart
RPL38	-	-	-	-	-	-	-	-	-	-	
COMT	-	T	-	-	T	-	-	-	-	-	T=Liver
RPS7	-	-	-	-	-	-	-	-	-	-	
HSPB1	-	-	-	-	-	-	-	-	-	T	T=Placenta

Table 11 Evaluation of Methods on testing dataset 1 and 2.

Summary of the errors for all method/dataset pairs with optimized parameters on 2 testing dataset. The maximum number of true positives (TP), true negatives (TN), False positives (FP) and false negatives (FN) is ten, with TP+FN=10 and TN+FP=10. The best possible scenario is TP=TN=10 where all true are found and all false are rejected. The best methods are marked in GREEN.

Method-Data	TP	TN	FP	FN
BioGPS-RS	6	10	0	4
BioGPS-DEC	5	9	1	5
GDS596-RS	2	10	0	8
GDS596-DEC	3	10	0	7
GDS596-BF	3	8	2	7
GeAZr-RS	8	10	0	2
GeAZr-DEC	9	10	0	1
GDS3113-RS	10	10	0	0
GDS3113-DEC	9	10	0	1
GDS3113-BF	8	7	3	2

4.4 Results summary

The selected methods are applied on the whole datasets using optimized parameters. The Scoring algorithm is used to combine the result from four datasets. To illustrate the specificity in the future discussion, we classified specific or 2-selective genes by significance using the coverage and total score $t_s(T)$. The specific genes and 2-selective genes are reported separately in

Table **12**.

The score for specific genes are marked in four levels:

1) Strong support with $t_s(T) = 1$;
2) High support with $0.8 \leq t_s(T) < 1$;
3) Medium-high support with $0.5 < t_s(T) < 0.8$;
4) Medium support with $t_s(T) = 0.5$.

The score for 2-selective genes are divided into two groups:

1) Strong support: $(t_s(T_1), t_s(T_2)) = (0.5, 0.5)$;
2) Medium support: $(t_s(T_1), t_s(T_2)) \geq (0.3, 0.3)$.

Please note that the criteria of score are tentative and can be changed by different interests of researches. Actually, as the methods that have been on datasets are in a strict manner and one can put a high confidence on the detected specific genes even with lower score (i.e. $t_s(T) < 0.5$) .

For a summary, there are 4554 specific genes and 598 2-selective genes detected under all coverage scope using the criteria that specificity greater than 0.5 and selectivity greater than 0.3.

Table 12 Summary of results. Total indicates the number of genes for each level of coverage.

Score / Coverage	Number of Specific Genes				Number of 2-Selective Genes		Total number of genes for each coverage
	1	0.8-1	0.5-0.8	0.5	0.5	>0.3	
4 of 4	117	148	510	592	6	39	9904
3 of 4	59	11	113	32	1	123	3768
2 of 4	91	3	63	503	39	3	2443
1 of 4	1688	0	10	6	382	5	11408
Summation	**1955**	**162**	**696**	**1133**	**428**	**170**	**27523**

4.5 Comparison with other databases

As shown in Table 12, there are 117 genes detected as specific with *strong support* ($t_s(T) = 1$), and 45 genes as 2-selective (6 with *strong support*, $t_s(T_1) = t_s(T_2) = 0.5$, and 39 with *medium-high* support $(t_s(T_1), t_s(T_2)) \geq (0.3, 0.3)$) under the coverage of 4. As these results are supported by all methods and are detectable by every datasets individually, which are very strict criterion, they are the best candidates of specific/selective genes that we can have. These genes are used for the evaluation of the result and are compared with the results from TiSGeD, TiGER and HPA. Table 13 shows the comparison of the 117 specific genes, and Table 14 shows the comparison of 45 2-selective genes.

Table 13 and 14 are colored by the degree of agreement between the predicted result and the databases. Those with GREY have exact match with our prediction, and those with RED have at least one matched tissue. It is easily captured by eye browsing that most of our predictions are supported at least one of these three databases. If we look at the performance of individual databases, TiSGeD has the most consistent results as we expected since they also use the microarray data; TiGER with EST data follows up; and HPA, which suffers a lot by missing expression data, has the least agreement with our prediction. Interestingly, there are cases where HPA and TiGER support our result while TiSGeD rejected. For example, NPHS2 is detected as ubiquitously expressed gene in TiSGeD while both HPA and TiGER report exact the same as our prediction with kidney-specific. The overlap for each database is shown in Figure 9 for 117 specific genes and Figure 10 for 45 2-selective genes. The fully agreement with our predictions is: 82% with TiSGeD, 75% with TiGER and only 30% with HPA (Fig. 2). And if we extend the proportion to partial match, the number goes up to 91% with TiSGeD, 84% with TiGER and 64% with HPA.

Similar to the specific genes, the overlap between our predicted results s and the results of the databases are shown in Figure 10. *Fully agree* means that both the predicted tissues must be exactly same. It is expected that this proportion (20% with TiGER, 31% with TiSGeD and 2% with HPA) is much lower than it for the 117 specific genes. But the percentage of *partially agree* (74% with TiGER, 91% with TiSGeD and 90% for HPA) which requires at least one same tissue is equally good as in the specific group.

Table 13:a Comparison of Specific Genes. (part 1)

Compare results of 117 specific genes with the best coverage and highest score to TiGER, TiSGeD and HPA. Grey shows the exact agreement in the corresponding database, red shows partially agreement, and absence and disagreement remain white. '/' means the gene is not found in the database and '-' means this gene is not specific.

	Predicted	TiGER	TiSGeD	HPA
FDXR	Adrenal	Cervix	Adrenal	Many tissues with strong, Adrenal (Strong)
CYP11B1	Adrenal	/	Adrenal	No expression data
HSD3B2	Adrenal	-	Adrenal	No expression data
DOCK3	CNS	-	CNS	Many tissues with strong, CNS (Strong)
CHN1	CNS	-	CNS	Several tissues with strong, CNS (moderate)
NRGN	CNS	CNS	CNS	Several tissues with strong, CNS (Strong)
CCK	CNS	Placenta	CNS	Several tissues with strong, CNS (Strong)
CACNG3	CNS	CNS	CNS	No expression data
LY6H	CNS	CNS	CNS	No expression data
RGS4	CNS	-	CNS, Heart	Many tissues with strong, CNS (Negative)
NELL2	CNS	Small intestine	Lung	Kidney (Strong)
TNNT2	Heart	Heart	Heart	Heart (Strong)
MYL7	Heart	Heart	Heart	Several tissues with strong, Heart (Strong)
MYBPC3	Heart	Heart	Heart	No expression data
SLC4A3	Heart	-	Heart	No expression data
SLC12A1	Kidney	Kidney	Kidney	Kidney (Strong)
NPHS2	Kidney	Kidney	-	Kidney (Strong)
KL	Kidney	Kidney	Kidney	Several tissues with strong, Kidney (Strong)
UMOD	Kidney	Kidney	Kidney	Several tissues with strong, Kidney (Strong)
CLCNKB	Kidney	Kidney	Kidney	No expression data
KCNJ1	Kidney	Kidney	Kidney	Kidney (Strong), Testis (Strong)
SLC12A3	Kidney	Cervix, Kidney	Kidney, Ovary	Several tissues with strong, Kidney (Moderate)
SLC34A1	Kidney	Kidney	Kidney	No expression data
CYP1A2	Liver	/	Liver	Liver (Strong)
CYP2A6	Liver	Liver	Liver	Liver (Strong)
F12	Liver	Liver, Stomach	Liver	No tissues with strong
SERPINA6	Liver	Liver	Liver	Several tissues with strong, Liver (Moderate)
SERPINF2	Liver	Liver, Kidney	Liver, Breast	Several tissues with strong, Liver (Strong)
SPP2	Liver	Liver	-	No expression data
SLC22A1	Liver	Liver	Liver	Uterine Cervix (Strong), Liver (Moderate)
C8G	Liver	Liver, Stomach	Liver	No expression data
HPR	Liver	Liver	Liver	No expression data
MAT1A	Liver	Liver	Liver	No expression data
ITIH1	Liver	Liver	Liver	No expression data
AGXT	Liver	Liver	Liver	Liver (Strong), Testis (Strong)
CYP2E1	Liver	Liver	Liver	Liver (Strong)
LIPC	Liver	Liver	Liver, Colon	No strong, Liver (Weak)
CYP2C8	Liver	Liver	Liver	Liver (Strong)
HP	Liver	Liver	-	No Strong, Liver (Weak)
MST1	Liver	Liver	Liver	Gall bladder (Strong), Liver (Moderate)

Table 10:b Comparison of Specific Genes. (part 2)

	Predicted	TiGER	TiSGeD	HPA
AGT	Liver	Liver	Liver	CNS (Strong), Placenta (Strong), Liver (Moderate)
APOC2	Liver	Liver	Liver	No expression data
SFTPC	Lung	Lung	Lung	Lung (Strong)
SFTPB	Lung	Lung	Lung	Lung (Moderate)
SFTPD	Lung	Lung	Lung	Lung (Strong)
AGER	Lung	Lung	Lung	Several tissues with strong, Lung (Strong)
IAPP	Pancreas	Pancreas	Pancreas	Pancreas (Strong)
PSG2	Placenta	Placenta	Placenta	No expression data
PSG7	Placenta	Placenta	Placenta, Lung	No expression data
PSG4	Placenta	Placenta	Placenta	No expression data
GCM1	Placenta	Placenta	Placenta	No expression data
PLAC1	Placenta	Placenta	Placenta, Kidney	No expression data
PSG11	Placenta	Placenta	Placenta	No expression data
PSG5	Placenta	Placenta	Placenta, Lung	No expression data
LGALS14	Placenta	Placenta	Placenta	No expression data
PSG6	Placenta	Placenta	Placenta	No expression data
PSG9	Placenta	Placenta	Placenta, Breast	No expression data
CAPN6	Placenta	Placenta	Placenta	Several tissues with strong, Placenta (Strong)
PSG3	Placenta	Placenta	Ovary	No expression data
CYP19A1	Placenta	Placenta	Placenta	Placenta (Strong)
FCGR2B	Placenta	Blood	Placenta	Appendix (Strong), Placenta (Moderate)
ADAM12	Placenta	Placenta	Placenta	Several tissues with strong, Placenta (Strong)
BMP1	Placenta	Placenta	Placenta	Several tissues with strong, Placenta (Strong)
HSD17B1	Placenta	Placenta	Placenta, Pancreas	Placenta (Strong)
CGB	Placenta	Placenta	Placenta, Osteosarcoma	Placenta (Strong), Epididymis (Strong)
EBI3	Placenta	Placenta	Placenta	Several tissues with strong, Testis (Moderate)
KLK2	Prostate	Prostate, Nerve	Prostate	Prostate (Strong)
ACPP	Prostate	Prostate	Prostate	Prostate (Strong)
CST4	Salivary gland	Bladder, Colon	Salivary gland	No expression data
TNNT3	Skeletal	Muscle	Skeletal muscle	Skeletal muscle (Strong)
CASQ1	Skeletal	Larynx, Muscle	Skeletal muscle	Several tissues with strong, Skeletal muscle (Strong)
ACTN3	Skeletal	Muscle	Skeletal muscle	/
PYGM	Skeletal	Muscle	Skeletal muscle	No expression data
FHL3	Skeletal	Muscle	Skeletal muscle	No expression data
TNNC2	Skeletal	Muscle	Skeletal muscle, Thyroid	No expression data
MYF6	Skeletal	Muscle, Heart	Skeletal muscle	No expression data
LOR	Skin	/	Skin	No expression data
AGTRL1	Spinal cord	Heart	/	CNS (Strong)

Table 10:c Comparison of Specific Genes. (part 3)

	Predicted	TiGER	TiSGeD	HPA
CDKN3	Testis	Bone	Testis	Epididymis (Strong), Placenta (Strong)
ADAM2	Testis	Testis	Testis	Testis (Strong)
INSL3	Testis	Brain	Testis	Testis (Strong)
ACTL7A	Testis	Testis	Testis	Testis (Moderate)
PRM2	Testis	Testis	Testis	Testis (Moderate)
AKAP4	Testis	Testis	Testis	Testis (Moderate)
ACTL7B	Testis	Testis	Testis	Testis (Strong)
TEX14	Testis	Testis	Testis	Many tissues with strong, Testis (Moderate)
C1orf14	Testis	Testis	Testis	Many tissues with strong, Testis (Strong)
BRDT	Testis	Testis	Testis	Many tissues with strong, Testis (Strong)
CCIN	Testis	Testis	Testis	Several tissues with strong, Testis (Strong)
APH1B	Testis	Testis	Testis	No expression data
CCNA1	Testis	Bone marrow, Testis	Testis	No expression data
OAZ3	Testis	Testis	Testis	No expression data
CABYR	Testis	Testis	Testis	No expression data
ODF1	Testis	/	Testis	No expression data
LDHC	Testis	Testis	Testis	No expression data
ANKRD7	Testis	Testis	Testis	No expression data
TNP1	Testis	Testis	Testis	No expression data
CCT6B	Testis	Testis	Testis	No expression data
DDX4	Testis	Testis	Testis	No expression data
C19orf36	Testis	Testis	/	No expression data
TCP11	Testis	Testis	Testis	No expression data
LOC81691	Testis	Testis	Testis	/
TPTE	Testis	Testis	Testis	No expression data
KCNK4	Testis	Testis	-	No expression data
PHF7	Testis	Testis	Testis	Testis (Strong)
HSPA1L	Testis	Testis	Testis	No expression data
SPA17	Testis	Testis	-	Several tissues with strong, Testis (Strong)
ODF2	Testis	-	Testis	No strong, Testis (Weak)
NXF3	Testis	Testis, Ovary, Stomach	Testis	Several tissues with strong, Testis (Strong)
PRSS16	Thymus	-	Thymus	Pancreas (Strong), Stomach (Strong)
CD1E	Thymus	Thymus	Thymus	No expression data
TCF7	Thymus	Thymus	-	Several tissues with strong, No Thymus
CD3D	Thymus	-	-	Several tissues with strong, No Thymus
DNTT	Thymus	/	-	No expression data
TPO	Thyroid	-	Thyroid	Thyroid (Strong)
SLC26A4	Thyroid	-	Thyroid	No expression data
TSHR	Thyroid	Ovary	Thyroid	Several tissues with strong, Thyroid (Strong)

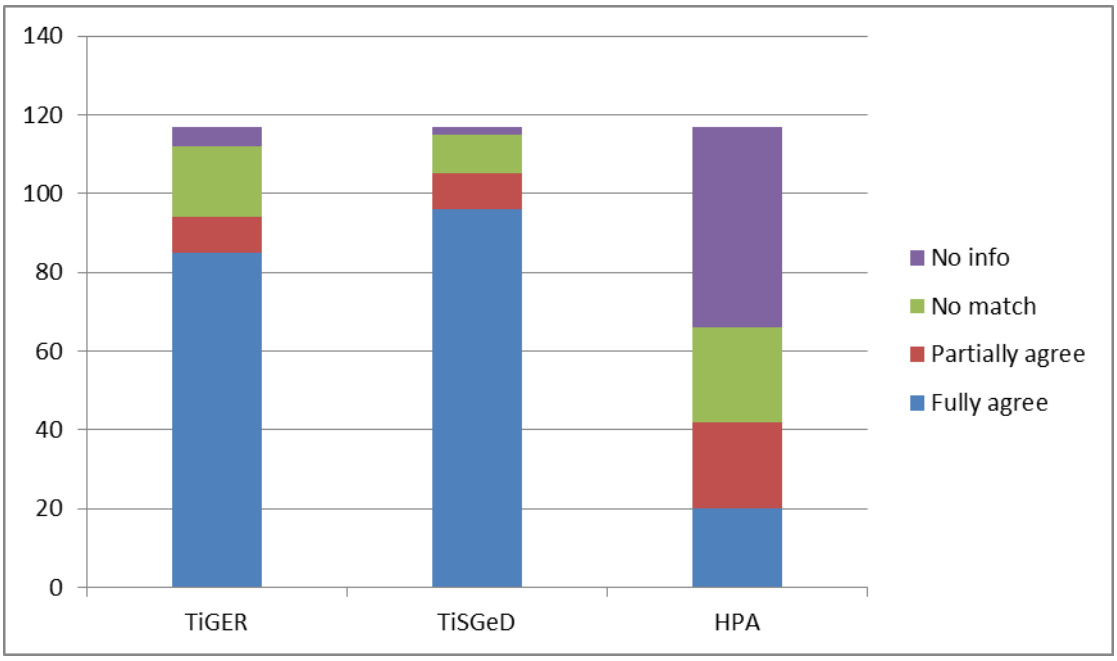

Figure 9 Comparison for 117 specific genes: histogram shows how exactly the other database agrees with the predicted result.

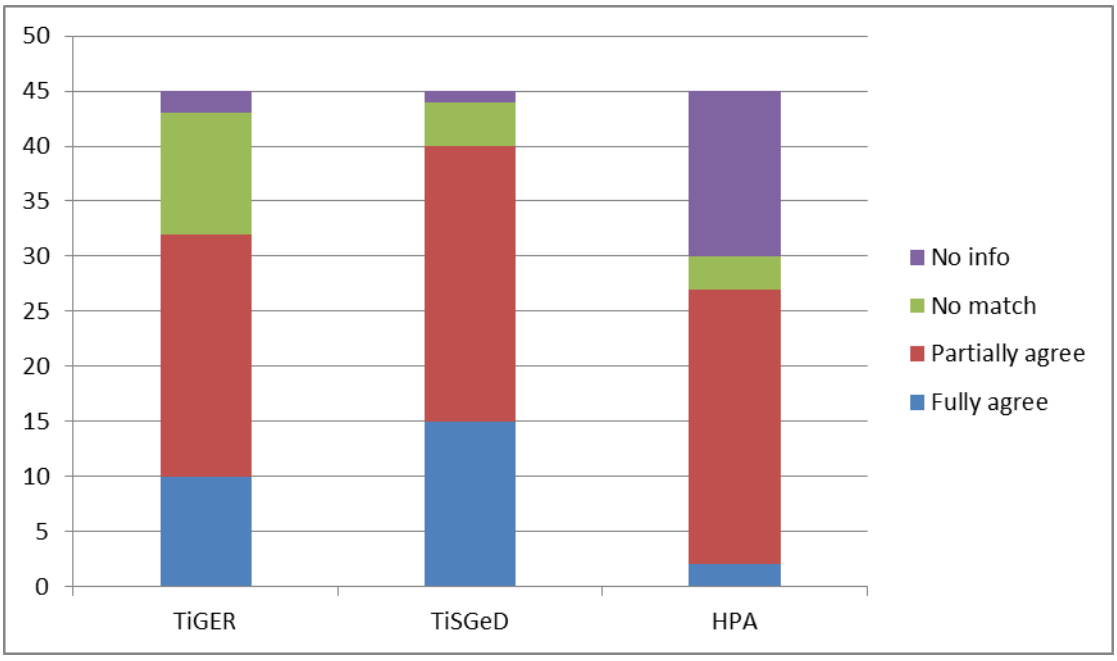

Figure 10 Comparison for 60 2-selective genes: Histogram shows how exactly the other database agrees with the predicted result.

Table 14:a Comparison of 2-selective Genes. (part 1)
Comparing the result of 6 2-selective genes with highest score (first six in the table) and 39 selective genes with medium score in the best coverage with TiGER, TiSGeD and HPA. Grey shows the exact agreement in the corresponding database, red shows partially agreement, and absence and disagreement remain white. '/' means the gene is not found in the database and '-' means this gene is not specific.

	PREDICTED TISSUES	TIGER	TISGED	HPA
UGT8	Spinal cord, CNS	-	Spinal cord, Colon, CNS	CNS (Strong)
SNAP25	Cerebellum, CNS	CNS	CNS	CNS (Strong), Pancreas (Strong)
TNNC1	Heart, Skeletal muscle	Heart, Muscle	Heart, Colon, Ovary	Heart (Strong),Skeletal muscle (Strong)
MYH7	Heart, Skeletal muscle	Heart, Muscle	Heart, Skeletal muscle	Heart (Strong), Skeletal muscle (Strong)
LRP2	Kidney, Thyroid	Kidney	Kidney, Thyroid	Heart (Strong), Kidney(Strong), Thyroid (Strong), Adrenal (Strong)
KNG1	Kidney, Liver	Kidney, Liver	Kidney Liver	Kidney (Strong)
DPYS	Kidney, Liver	Kidney, Liver	Kidney, Liver	Kidney (Strong), Liver (Moderate)
RGN	Liver, Adrenal	Kidney	Liver, Adrenal	Liver (Strong), Adrenal (Strong), Testis (Strong)
MYOM2	Heart, Skeletal muscle	Heart, Muscle	Heart	Many tissues with strong, Heart (Strong), Skeletal Muscle (Strong)
PVALB	Cerebellum, CNS	Kidney	CNS, Cerebellum	Parathyroid gland (Strong), Cerebellum (Strong)
CGA	Placenta, CNS	Placenta	Placenta,CNS, Salivary gland	Placenta (Strong)
GH1	Placenta, CNS	CNS	-	Placenta (moderate)
CSF3R	Blood, Placenta	Blood, Placenta	Blood, Placenta	Placenta (Strong)
LPO	Salivary gland, Trachea	Blood, Muscle	Salivary gland, Trachea	Salivary gland (Strong)
MYL3	Heart, Skeletal muscle	Heart, Muscle	Heart	Several tissues with strong, Heart (Strong), Skeletal Muscle (Strong)
MYL2	Heart, Skeletal muscle	Heart, Muscle	Heart	Several tissues with strong, Heart (Strong), Skeletal Muscle (Strong)
FBXO40	Heart, Skeletal muscle	-	Skeletal muscle	Several tissues with strong, Heart (Moderate), Skeletal muscle (Moderate)
ENPEP	Kidney, Small intestine	Kidney	Kidney	Several tissues with strong, Kidney (Strong), Small intestine (Strong)
HPD	Kidney, Liver	Liver	Liver	Several tissues with strong, Liver (Strong)

Table 13:b Comparison of 2-selective Genes. (part 2)

	PREDICTED TISSUES	TIGER	TISGED	HPA
MSMB	Prostate, Trachea	Prostate	Prostate, Trachea	Stomach (Strong), Bronchus (Strong), Prostate (Strong)
ART3	Skeletal muscle, Testis	Muscle, Testis	Testis, Breast	Testis(Strong)
CLGN	Heart, Testis	Testis	Testis	Testis (Strong), Fallopian tube (Strong)
SPRR1A	Tongue, Tonsil	Larynx, Tongue	Colon, Tongue, Thymus, Trachea	All Strong
PGAM2	Heart, Skeletal muscle	Muscle	Heart	No expression data
GLYAT	Kidney, Liver	Kidney, Liver	Kidney, Liver	No expression data
TM4SF5	Liver, Small intestine	Kidney	Liver	No expression data
REG3A	Pancreas, Small intestine	Stomach	Pancreas	No expression data
CCL20	Tonsil, Liver	Colon	Tonsil	No expression data
PGLYRP1	Bone marrow, Blood	/	Bone marrow, Pancreas	No expression data
CLEC4M	Liver, Lymph node	Placenta	Nerve	No expression data
PCSK1	Pancreas, CNS	Pancreas	Pancreas, CNS	No expression data
HSD3B1	Placenta, Adrenal	Placenta	Placenta, Adrenal	No expression data
AKR7A3	Liver, Small intestine	Colon, Stomach	Liver	No expression data
MOG	CNS, Spinal cord	CNS	CNS, Spinal cord	No Strong
IGFBP1	Placenta, Liver	Placenta	Placenta	Placenta (Strong), Liver (Weak)
CLC	Bone marrow, Blood	Bone marrow	Bone marrow	No expression data
PRB4	Trachea, Salivary gland	/	/	No expression data
EDN3	Retina, Salivary gland	Pancreas	Salivary gland, CNS	No expression data
SULT2A1	Liver, Adrenal	Liver	Liver, Adrenal	Several tissues with strong, Liver (Strong), Adrenal (Strong)
TFDP2	Thymus, Testis	-	-	Many tissues with strong, Testis (Strong)
RYR2	Heart, CNS	Heart	Heart	Many tissues with strong, CNS (Strong), Heart (Moderate)
KHK	Liver, Kidney	Liver	Liver	Liver (Strong), Kidney (Strong), Small intestine (Strong)
GABRD	Cerebellum, CNS	CNS	CNS, Cerebellum	No expression data
APOM	Liver, Kidney	Liver	Liver	No expression data
CES2	Liver, Small intestine	Liver	-	Many tissues with Strong, Liver (Strong), Small intestine (Strong)

4.6 Specificity analysis from a tissue aspect

We also analyzed which tissues are most frequently detected among the 1955 specific genes with *strong support* ($t_s(T) = 1$), and 598 2-selective genes with both *strong* and *medium support*.

The top 10 tissues (or tissue pairs) are shown in Figure 11 for specific gene and Figure 12 for 2-selective genes. For specific genes, the testis is with about 25% the top candidate, followed by CNS, Bile, Trachea, Blood, Placenta, Liver, Salivary gland, Heart and Skeletal muscle.

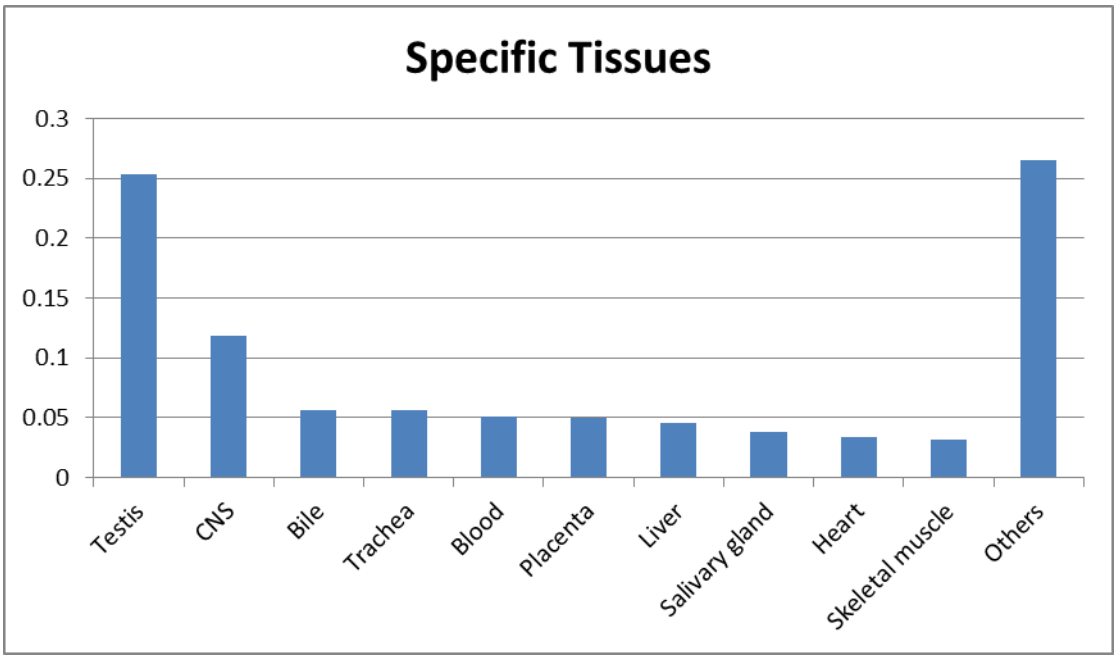

Figure 11 Top 10 most frequently occurred tissue among 1955 specific genes.

To investigate the result from a tissue aspect, Table 15 also provides answers to the question that which genes are specifically expressed in a certain tissue. For example, there are 88 genes that are specific in liver with *Strong* support. The statistic in table 15, the answer are restrict to score=1 regardless the value of coverage, while for the further study, researchers can define their interest on self-defined constraints on the score and coverage.

Table 15 The frequency of detected tissues among the 1955 specific genes.

Tissue name	Frequency	Relative frequency
Testis	494	0.2537
CNS	230	0.1181
Bile	110	0.0565
Trachea	110	0.0565
Blood	100	0.0514
Placenta	98	0.0503
Liver	88	0.0452
Salivary gland	74	0.0380
Heart	65	0.0334
Skeletal muscle	62	0.0318
Meniscus of joint	52	0.0267
Kidney	47	0.0241
Thymus	40	0.0205
Vessel	31	0.0159
Pancreas	29	0.0149
Skin	28	0.0144
Adrenal	27	0.0139
Epididymis	25	0.0128
Cerebellum	24	0.0123
Articular surface of bone	23	0.0118
Small intestine	22	0.0113
Lung	21	0.0108
Tonsil	20	0.0103
Prostate	19	0.0098
Thyroid	15	0.0077
Fallopian tube	9	0.0046
Retina	9	0.0046
Smooth muscle	8	0.0041
Spleen	7	0.0036
Bone marrow	7	0.0036
UHR	6	0.0031
Bone structure	6	0.0031
Ovary	6	0.0031
Spinal cord	5	0.0026
Stomach	5	0.0026
Endometrium	4	0.0021
Colon	3	0.0015
Tongue	3	0.0015
Uterus	3	0.0015
Brest	2	0.0010
Soft tissue	2	0.0010
Urethra	2	0.0010
Vas deferens	1	0.0005
Ureter	1	0.0005
Esophagus	1	0.0005
Tendon	1	0.0005
Vulva	1	0.0005

Interestingly, a large proportion of frequently detected tissue pairs (Table 16) seem to be functionally related. For example, the CNS (Central Nerve System) and spinal cord are both part of the nerve system; the skeletal muscle is similar to heart muscle; in some cases, the cerebellum can even be part of CNS; etc. This phenomenon can be discussed in various perspectives (See Section 5).

Table 16 The frequency of pairs of detected tissues among the 598 2-selective genes. Only those with frequencies above five are shown.

Tissue pairs	Frequency	Relative Frequency
CNS & Spinal cord	59	0.0983
Heart & Skeletal muscle	36	0.0600
CNS & Cerebellum	26	0.0433
Testis & Trachea	19	0.0317
Kidney & Liver	16	0.0267
Blood & Bone marrow	15	0.0250
Meniscus of joint & Nerve	13	0.0217
Liver & Small intestine	11	0.0183
Skin & Vulva	11	0.0183
Bile & Nerve	10	0.0167
Skin & Tonsil	10	0.0167
Blood & UHR	9	0.0150
Salivary gland & Trachea	9	0.0150
Testis & UHR	8	0.0133
Bile & Pancreas	8	0.0133
CNS & Testis	8	0.0133
CNS & Nerve	7	0.0117
CNS & Placenta	7	0.0117
Nerve & Pineal	6	0.0100
Bile & Salivary gland	6	0.0100

There are 195 tissue pairs have been identified among the 598 2–selective genes, while only 48 specific tissues in 1955 specific genes. The primary pair, CNS and spinal cord which only take about 10% of the total frequency, (Figure 12) can still be account for a significant proportion among all tissues pairs, and the top 10 frequent pairs takes almost 40% out of total. The ‘Other’ (61%) is the summation of the 185 tissue pairs apart from the top 10 pairs.

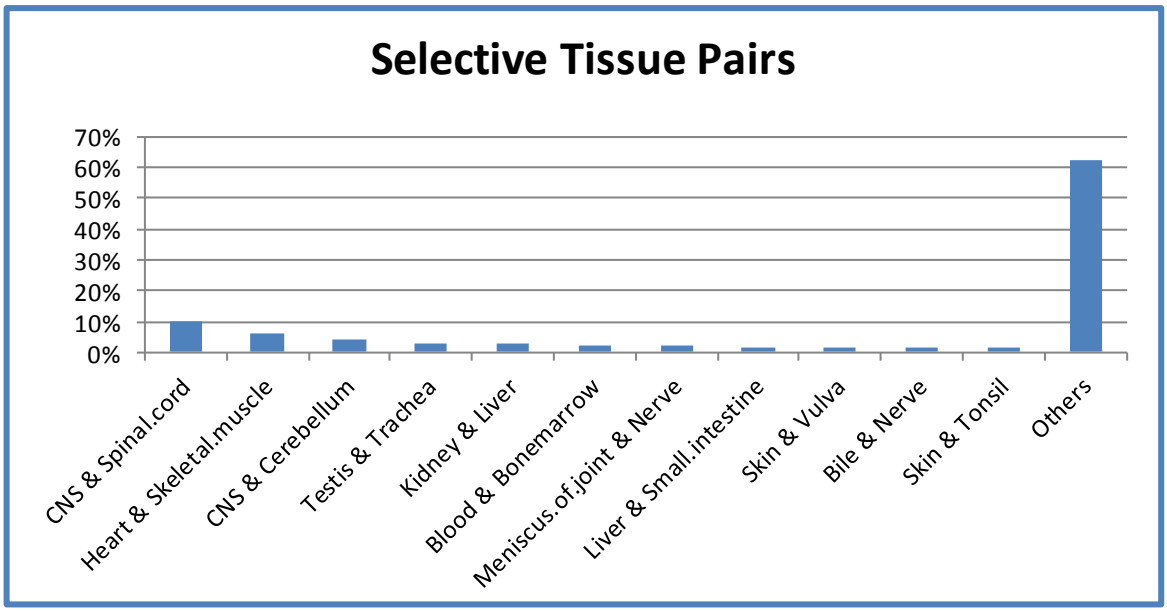

Figure 12 Top 10 most frequently occurred tissue pairs among 598 2-selective genes.

5 Conclusion and discussion

This project integrates the information of specificity from four large-scale datasets using the optimized methods after comparative analysis. The remarkable and unique features of this approach are the use of optimized method, the integration of different datasets, and the confidence indicator (score and coverage). The integrate result is compared with three external datasets which represent three types of expression data: the data from same kind (TiSGeD with microarray data), the data from similar technology (TiGER with EST data), and the proteomic data (HPA with antibody). The comparison confirmed the preciseness of the result, the advantage of comprehensive analysis, and the reliability of the prediction.

During the optimization of the methods, a set of tissue-specific genes are used to evaluate the parameters (Section 3.2). One may argue that the data would be over-fitted by training on only specific genes. However, two sets of testing genes (testing set 2) are used to verify the quality of the method in a later stage, and a ubiquitously expressed gene set is included. The result in Table 10 confirmed that the ubiquitously expressed genes can be properly detected.

To investigate the integrated results in Table 12, two parameters, the ‘coverage’ and the ‘score’, assess the reliability of the results. Obviously, genes with higher scores or coverage are more likely to be specific than those with lower values. However, it is more complicated to analyze genes with low support but high coverage or genes with high support but low coverage. The coverage is the existence of a particular gene in the four datasets, which is largely dependent on the enrichment of data sources. For example, GeAZr data almost doubled the number of genes in BioGPS and GDS596. Genes with ‘coverage one’ are mostly from GeAZr data, or the probesets without annotated gene names. The coverage can be considered as an auxiliary mark. The score, which is the primary indicator of specificity, can be affected by the tissue vocabulary of each dataset and the quality of probesets for each gene. For instance, the tissue stomach is only included by GeAZr and GDS3113, the lack of this item in BioGPS and GDS596 can be the cause of low score for stomach specificity in genes like LIPF. On the other hand, as there are probesets of bad quality, the average algorithm, which treats each probeset equally, may produce noise to the final score. For example, if three out of five probesets of a same gene are specific in heart, and the other two, which are actually of bad quality, are specific in liver, the algorithm will give heart 0.6 instead of full score. In order to diminish the noisy level, we have improved the scoring algorithm by introducing the majority vote in the probesets (Section 3.4.3); however, as there is no authority for the quality of probeset, it is difficult to give an exact score in large scale level. The scoring algorithm in this project is experimental and aims to solve the problem in a simple manner. For the future work, methods using statistical model or machine learning approaches might be inspiring.

The co-occurring 2-selective tissue pairs with related functions are not rare cases in Section 4.4 (Table 17). To some extent, this has confirmed that gene is expected to express in functionally related tissues. This point gives a suggestion to the future work:

1. Examine the expression pattern among cell lines instead of separate tissues [42-43]. A gene that is functionally important to a particular cell type can be expressed across many tissues that contain this type of cells. In this case, the gene with a biological meaning of specificity probably cannot be detected as specific in any tissue. The detection of cell line specific genes might be a valuable complement in both functional biology and pharmaceutical development.
2. To another extend of interest, the result of functional similarity of tissue pairs, is also an issue of vocabulary mapping. For example, should the CNS and cerebellum be grouped together or not? According to the principle of this project, since cerebellum occurs in three datasets out of four, it is more reasonable to keep it as an independent term. Contradictory, from the biological point of view, the cerebellum is a part of central nerve system, and in this case, these 2-selective genes are actually CNS-specific genes.

The integration of result is basically achieved by averaging the results over four different dataset/method pairs. However, instead of this computational meta-analysis, an alternative approach can be started from integrating the microarray data in the probeset level before applying the detection methods. The advantage of this approach is obvious: more reliable expression data save of computational resources; no need of multiple methods; avoidance of the bias caused by the combination of results. This approach requires more expertise on large scale transcriptomic technology and profound biological knowledge. Further, it is not evident how to combine data obtained from different platforms.

From the technologies perspective, RNA-Seq, a recently developed large-scale profiling method using deep-sequencing technology can provide a more precise, more efficient and less expensive measurement on the trancriptomic level [44]. A few studies have been performed on this new data type: an analysis across human and mouse tissue and cell lines reported approximately 8000 ubiquitously expressed genes [45]; another research on soybean trancriptome identified the most highly expressed and the legumes-specific genes [46], etc. Of course, as a technology in its early developing stage, RNA-seq has challenges to be conquered, such as the enrichment of data, the target of more complex trancriptomes, ect. However, it is expectable that this kind of new technology will play the main role in the future trancriptomes and provide significant benefits to the study of gene expressional specificity.

In conclusion, this study can provide meaningful indication for the prediction of innovative drug targets, and a valuable reference to the other technologies of expression-pattern-relevant studies.

References

1. Yang, Y., S.J. Adelstein, and A.I. Kassis, *Target discovery from data mining approaches.* Drug Discov Today, 2009. **14**(3-4): p. 147-54.
2. Van Deun, K., et al., *Testing the hypothesis of tissue selectivity: the intersection-union test and a Bayesian approach.* Bioinformatics, 2009. **25**(19): p. 2588-94.
3. Klee, E.W., *Data mining for biomarker development: a review of tissue specificity analysis.* Clin Lab Med, 2008. **28**(1): p. 127-43, viii.
4. Greller, L.D. and F.L. Tobin, *Detecting selective expression of genes and proteins.* Genome Res, 1999. **9**(3): p. 282-96.
5. Dezso, Z., et al., *A comprehensive functional analysis of tissue specificity of human gene expression.* BMC Biol, 2008. **6**: p. 49.
6. Terstappen, G.C. and A. Reggiani, *In silico research in drug discovery.* Trends Pharmacol Sci, 2001. **22**(1): p. 23-6.
7. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.
8. Young, R.A., *Biomedical discovery with DNA arrays.* Cell, 2000. **102**(1): p. 9-15.
9. Brown, P.O. and D. Botstein, *Exploring the new world of the genome with DNA microarrays.* Nat Genet, 1999. **21**(1 Suppl): p. 33-7.
10. Lipshutz, R.J., et al., *High density synthetic oligonucleotide arrays.* Nat Genet, 1999. **21**(1 Suppl): p. 20-4.
11. Barrett, T., et al., *NCBI GEO: mining tens of millions of expression profiles--database and tools update.* Nucleic Acids Res, 2007. **35**(Database issue): p. D760-5.
12. Wu, C., et al., *BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources.* Genome Biol, 2009. **10**(11): p. R130.
13. Parkinson, H., et al., *ArrayExpress--a public database of microarray experiments and gene expression profiles.* Nucleic Acids Res, 2007. **35**(Database issue): p. D747-50.
14. Whitehead, A. and D.L. Crawford, *Variation in tissue-specific gene expression among natural populations.* Genome Biol, 2005. **6**(2): p. R13.
15. Hsiao, L.L., et al., *A compendium of gene expression in normal human tissues.* Physiol Genomics, 2001. **7**(2): p. 97-104.
16. Haverty, P.M., et al., *HugeIndex: a database with visualization tools for high-density oligonucleotide array data from normal human tissues.* Nucleic Acids Res, 2002. **30**(1): p. 214-7.
17. Kadota, K., et al., *Detection of genes with tissue-specific expression patterns using Akaike's information criterion procedure.* Physiol Genomics, 2003. **12**(3): p. 251-9.
18. Ge, X., et al., *Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues.* Genomics, 2005. **86**(2): p. 127-41.
19. Kadota, K., et al., *ROKU: a novel method for identification of tissue-specific genes.* BMC Bioinformatics, 2006. **7**: p. 294.
20. Tuke, J., G.F.V. Glonek, and P.J. Solomon, *Gene profiling for determining pluripotent genes in a time course microarray experiment.* Biostatistics, 2009. **10**(1): p. 80-93.
21. Adams, M.D., et al., *Complementary DNA sequencing: expressed sequence tags and human genome project.* Science, 1991. **252**(5013): p. 1651-6.

22. Skrabanek, L. and F. Campagne, *TissueInfo: high-throughput identification of tissue expression profiles and specificity.* Nucleic Acids Res, 2001. **29**(21): p. E102-2.
23. Brown, A.C., et al., *ExQuest, a novel method for displaying quantitative gene expression from ESTs.* Genomics, 2004. **83**(3): p. 528-39.
24. Velculescu, V.E., et al., *Serial analysis of gene expression.* Science, 1995. **270**(5235): p. 484-7.
25. Jongeneel, C.V., et al., *An atlas of human gene expression from massively parallel signature sequencing (MPSS).* Genome Res, 2005. **15**(7): p. 1007-14.
26. Kouadjo, K.E., et al., *Housekeeping and tissue-specific genes in mouse tissues.* BMC Genomics, 2007. **8**: p. 127.
27. Guo, M., et al., *Genome-wide allele-specific expression analysis using Massively Parallel Signature Sequencing (MPSS) reveals cis- and trans-effects on gene expression in maize hybrid meristem tissue.* Plant Mol Biol, 2008. **66**(5): p. 551-63.
28. Boon, K., et al., *An anatomy of normal and malignant gene expression.* Proc Natl Acad Sci U S A, 2002. **99**(17): p. 11287-92.
29. Schug, J., et al., *Promoter features related to tissue specificity as measured by Shannon entropy.* Genome Biol, 2005. **6**(4): p. R33.
30. Reverter, A., A. Ingham, and B.P. Dalrymple, *Mining tissue specificity, gene connectivity and disease association to reveal a set of genes that modify the action of disease causing genes.* BioData Min, 2008. **1**(1): p. 8.
31. Watson JD, H.N., Roberts JW, Steitz JA, Weiner AM, *The functioning of higher eucaryotic genes,* in *Molecular biology of the gene* 1965. p. 704.
32. *BioGPS*. Available from: http://biogps.gnf.org/downloads/.
33. Su, A.I., et al., *Large-scale analysis of the human and mouse transcriptomes.* Proc Natl Acad Sci U S A, 2002. **99**(7): p. 4465-70.
34. Xiao, S.J., et al., *TiSGeD: a database for tissue-specific genes.* Bioinformatics, 2010. **26**(9): p. 1273-5.
35. Liu, X., et al., *TiGER: a database for tissue-specific gene expression and regulation.* BMC Bioinformatics, 2008. **9**: p. 271.
36. Uhlen, M., et al., *A human protein atlas for normal and cancer tissues based on antibody proteomics.* Mol Cell Proteomics, 2005. **4**(12): p. 1920-32.
37. Ueda, T., *A simple method for the detection of outliers.* Electronic Journal of Applied Statistical Analysis, 2009(1): p. 67-76.
38. Kitagawa, G., *Use of Aic for the Detection of Outliers.* Technometrics, 1979. **21**(2): p. 193-199.
39. Kadota, K., T. Konishi, and K. Shimizu, *Evaluation of two outlier-detection-based methods for detecting tissue-selective genes from microarray data.* Gene Regul Syst Bio, 2007. **1**: p. 9-15.
40. Deun, K.V. *BayesianIUT Software*. 2009; Available from: [http://ppw.kuleuven.be/okp/software/bayesianiut/].
41. *R project*. R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences,

but much code written for S runs unaltered under R.]. Available from: http://www.r-project.org/.

42. Nelander, S., P. Mostad, and P. Lindahl, *Prediction of cell type-specific gene modules: identification and initial characterization of a core set of smooth muscle-specific genes.* Genome Res, 2003. **13**(8): p. 1838-54.
43. Nelander, S., et al., *Predictive screening for regulators of conserved functional gene modules (gene batteries) in mammals.* BMC Genomics, 2005. **6**: p. 68.
44. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
45. Ramskold, D., et al., *An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data.* PLoS Comput Biol, 2009. **5**(12): p. e1000598.
46. Severin, A.J., et al., *RNA-Seq Atlas of Glycine max: a guide to the soybean transcriptome.* BMC Plant Biol, 2010. **10**: p. 160.

Appendix

A - 1: The vocabulary mapping for BioGPS.

New Term	Original Term	New Term	Original Term
Testis	Testis	**Bone.marrow**	Bonemarrow
	TestisGermCell	**Kidney**	Kidney
	TestisIntersitial	**Ovary**	Ovary
	TestisLeydigCell	**Placenta**	Placenta
	TestisSeminiferousTubule	**Prostate**	Prostate
Uterus	Uterus	**Salivary.gland**	Salivarygland
	UterusCorpus	**Skin**	Skin
Pineal	pineal_day	**Spinal.cord**	Spinalcord
	pineal_night	**Thymus**	Thymus
Pancreas	Pancreas	**Tongue**	Tongue
	PancreaticIslet	**Tonsil**	Tonsil
Skeletal.muscle	SkeletalMuscle	**Trachea**	Trachea
Smooth.muscle	SmoothMuscle	**Blood**	WholeBlood
Cardiac.myocytes	CardiacMyocytes	**Colon**	colon
Adrenal	AdrenalCortex	**Retina**	retina
	Adrenalgland	**Small.intestine**	small_intestine
Cerebellum	Cerebellum	**Thyroid**	Thyroid
	CerebellumPeduncles	**Liver**	Liver
CNS	Caudatenucleus	**Lung**	Lung
	GlobusPallidus	**Lymph.node**	Lymphnode
	MedullaOblongata	**REMOVED**	CD105+_Endothelial
	OlfactoryBulb		CD14+_Monocytes
	Pituitary		CD19+_BCells(neg._sel.)
	Pons		CD33+_Myeloid
	SubthalamicNucleus		CD34+
	Hypothalamus		CD4+_Tcells
	Thalamus		CD56+_NKCells
	Wholebrain		CD71+_EarlyErythroid
	Amygdala		CD8+_Tcells
	OlfactoryBulb		BronchialEpithelialCells
	ParietalLobe		BDCA4+_DentriticCells
	TemporalLobe		FetalThyroid
	CingulateCortex		Fetalbrain
	PrefrontalCortex		Fetalliver
Nerve	CiliaryGanglion		Fetallung
	DorsalRootGanglion		Leukemia_chronicMyelogenousK-562
	SuperiorCervicalGanglion		Leukemia_promyelocytic-HL-60
	TrigeminalGanglion		Leukemialymphoblastic(MOLT-4)
Heart	Heart		Lymphoma_burkitts(Daudi)
	AtrioventricularNode		Lymphoma_burkitts(Raji)
Adipocyte	Adipocyte		721_B_lymphoblasts
Appendix	Appendix		Colorectaladenocarcinoma

A-2: The vocabulary mapping for GDS596

New Term	Original Term	New Term	Original Term
CNS	Caudatenucleus	**Bonemarrow**	Bonemarrow
	GlobusPallidus	**Kidney**	Kidney
	MedullaOblongata	**Ovary**	Ovary
	OlfactoryBulb	**Placenta**	Placenta
	Pituitary	**Prostate**	Prostate
	Pons	**Salivary.gland**	Salivarygland
	SubthalamicNucleus	**Skin**	Skin
	Hypothalamus	**Spinal.cord**	Spinalcord
	Thalamus	**Thymus**	Thymus
	Wholebrain	**Tongue**	Tongue
	Amygdala	**Tonsil**	Tonsil
	Occipital lobe	**Trachea**	Trachea
	ParietalLobe	**Blood**	WholeBlood
	TemporalLobe	**Thyroid**	Thyroid
	CingulateCortex	**Liver**	Liver
	PrefrontalCortex	**Lung**	Lung
Testis	Testis	**Lymph.node**	Lymphnode
	TestisGermCell	REMOVED	CD105+_Endothelial
	TestisIntersitial		CD14+_Monocytes
	TestisLeydigCell		CD19+_BCells(neg._sel.)
	TestisSeminiferousTubule		CD33+_Myeloid
Uterus	Uterus		CD34+
	UterusCorpus		CD4+_Tcells
Pancreas	Pancreas		CD56+_NKCells
	PancreaticIslet		CD71+_EarlyErythroid
Skeletal.muscle	SkeletalMuscle		CD8+_Tcells
Smooth.muscle	SmoothMuscle		BronchialEpithelialCells
Cardiac.myocytes	CardiacMyocytes		BDCA4+_DentriticCells
Adrenal	AdrenalCortex		FetalThyroid
	Adrenalgland		Fetalbrain
Cerebellum	Cerebellum		Fetalliver
	CerebellumPeduncles		Fetallung
Nerve	CiliaryGanglion		Leukemia_chronicMyelogenousK-56
	DorsalRootGanglion		Leukemia_promyelocytic-HL-60
	SuperiorCervicalGanglion		Leukemialymphoblastic(MOLT-4)
	TrigeminalGanglion		Lymphoma_burkitts(Daudi)
Heart	Heart		Lymphoma_burkitts(Raji)
	AtrioventricularNode		721_B_lymphoblasts
Adipocyte	Adipocyte		Colorectaladenocarcinoma
Appendix	Appendix		

A-3: The vocabulary mapping for GeAZr (part 1)

New Term	Original Term	New Term	Original Term
Vessel	Aorta	**Epididymis**	Epididymis
	Artery	**Prostate**	Prostate
	Abdominal aorta	**Seminal vesicle**	Seminal vesicle
	Ascending aorta	**Testis**	Testis
	Blood vessel	**Vas deferens**	Vas deferens
	Coronary artery	**Articular surface of bone**	Articular surface of bone
	Vein	**Bone structure**	Bone structure
Heart	Heart	**Meniscus of joint**	Meniscus of joint
	Left atrium	**Tendon**	Tendon and tendon sheath
	Left ventricle	**Soft tissue**	Soft tissues
	Right atrium	**Omentum**	Omentum
	Right ventricle	**Skeletal muscle**	Skeletal muscle
Bile	Common bile duct	**Smooth muscle**	Smooth muscle
	Gallbladder	**CNS**	Brain
Liver	Hepatic duct		Frontal cortex
	Liver		Occipital cortex
Pancrea	Pancreas		Temporal cortex
Salivary gland	Parotid gland		Parietal cortex
	Salivary gland		Corpus callosum
Appendix	Appendix		Locus ceruleus
Colon	Colon		Caudate nucleus
Small intestine	Duodenum		Putamen
	Ileum		Nucleus Accumbens
	Jejunum		Globus pallidus
	Small intestine		Subthalamic nucleus
Esophagus	Esophagus		Substantia nigra
Rectum	Rectum		Medulla oblongata
Stomach	Stomach		Nucleus basalis of Meynert
Tongue	Tongue		Amygdaloid nucleus
Adrenal	Adrenal cortex		Hippocampus
	Adrenal gland		Hypothalamus
Thyroid	Thyroid gland		Pulvinar
Cervix	Cervix		Thalamus
Endometrium	Endometrium		Cingulate gyrus
Fallopian tube	Fallopian tube		Entorhinal cortex
Myometrium	Myometrium		Pituitary gland
Ovary	Ovary		Red nucleus
Uterus	Uterus		White matter of occipital lobe

A-3: The vocabulary mapping for GeAZr (part 2)

New Term	Original Term	New Term	Original Term
Vagina	Vagina	**Cerebellum**	Cerebellum
Vulva	Vulva	**Nerve**	Nerve
Lymph node	Lymph node	**Spinal cord**	Spinal cord
Spleen	Spleen	**Placenta**	Placenta
Thymus	Thymus	**Bronchus**	Bronchus
Tonsil	Tonsil	**Larynx**	Larynx
Blood	White blood cell	**Lung**	Lung
Adipocyte	Perirenal fat	**Trachea**	Trachea
	Adipose tissue	**Bladder**	Bladder
	Adipose tissue of breast	**Kidney**	Kidney
Brest	Breast	**Ureter**	Ureter
Skin	Skin	**Urethra**	Urethra

A-3: The vocabulary mapping for GDS3113

New Term	Original Term
Lung	lung
Liver	liver
UHR	Universal.Human.Reference
CNS	brain
Prostate	prostate
Skeletal.muscle	skeletal.muscle
Heart	heart
Spinal.cord	spinal.cord
Tonsil	tonsil
Trachea	trachea
Uterus	uterus
Small.intestine	small.intestine
Skin	skin
Ovary	ovary
Testis	testis
Pancreas	pancreas
Thymus	thymus
Kidney	kidney
Placenta	placenta
Thyroid	thyroid
Salivary.gland	salivary.gland
Colon	colon
Mammary.gland	mammary.gland
Spleen	spleen
Adrenal	adrenal.gland
Blood	peripheral.blood.lymphocyte
Bonemarrow	bone.marrow
Retina	retina
REMOVED	fetal.liver
	fetal.brain
	fetal.thymus
	fetal.kidney